

A Note on Spectral Clustering

Pavel Kolev* Kurt Mehlhorn
 Max-Planck-Institut für Informatik, Saarbrücken, Germany
 {pkolev,mehlhorn}@mpi-inf.mpg.de

Abstract

Spectral clustering is a popular and successful approach for partitioning the nodes of a graph into clusters for which the ratio of outside connections compared to the volume (sum of degrees) is small. In order to partition into k clusters, one first computes an approximation of the first k eigenvectors of the (normalized) Laplacian of G , uses it to embed the vertices of G into k -dimensional Euclidean space \mathbb{R}^k , and then partitions the resulting points via a k -means clustering algorithm. It is an important task for theory to explain the success of spectral clustering.

Peng et al. (COLT, 2015) made an important step in this direction. They showed that spectral clustering provably works if the gap between the $(k+1)$ -th and the k -th eigenvalue of the normalized Laplacian is sufficiently large. They prove a structural and an algorithmic result. The algorithmic result needs a considerably stronger gap assumption and does not analyze the standard spectral clustering paradigm; it replaces spectral embedding by heat kernel embedding and k -means clustering by locality sensitive hashing.

We extend their work in two directions. Structurally, we improve the quality guarantee for spectral clustering by a factor of k and simultaneously weaken the gap assumption. Algorithmically, we show that the standard paradigm for spectral clustering works. Moreover, it even works with the same gap assumption as required for the structural result.

*This work has been funded by the Cluster of Excellence “Multimodal Computing and Interaction” within the Excellence Initiative of the German Federal Government.

Contents

1	Introduction	2
2	Technical Contributions and Structure of the Paper	4
2.1	Exact Spectral Embedding - Notation	4
2.2	Exact Spectral Embedding - Our Results	5
2.3	Approximate Spectral Embedding - Notation	7
2.4	Approximate Spectral Embedding - Our Results	8
3	The Proof of Part (a) of The Main Theorem:	9
4	Vectors \hat{g}_i and f_i are Close	10
4.1	Analyzing the Columns of Matrix \mathbf{F}	10
4.2	Analyzing Eigenvectors f in terms of \hat{f}_j	12
5	Spectral Properties of Matrix \mathbf{B}	13
5.1	Analyzing the Column Space of Matrix \mathbf{B}	13
5.2	Analyzing the Row Space of Matrix \mathbf{B}	14
6	Proof of Lemma 2.4	17
7	The Normalized Spectral Embedding is ε-separated	19
8	An Efficient Spectral Clustering Algorithm	21
8.1	Proof of Lemma 2.7	21
8.2	Proof of Theorem 2.8	22
8.3	Proof of Theorem 2.9	24
8.4	Proof of Part (b) of Theorem 1.2	26
9	Parameterized Upper Bound on $\hat{\rho}_{\text{avr}}(k)$	27
A	Singular Values Bounds of Random Matrices	31

1 Introduction

A *cluster* in an undirected graph $G = (V, E)$ is a set S of nodes whose volume is large compared to the number of outside connections. Formally, we define the *conductance* of S by $\phi(S) = |E(S, \bar{S})| / \mu(S)$, where $\mu(S) = \sum_{v \in S} \deg(v)$ is the *volume* of S . The k -way partitioning problem for graphs asks to partition the vertices of a graph such that the conductance of each block of the partition is small (formal definition below). This problem arises in many applications, e.g., image segmentation and exploratory data analysis. We refer to the survey [12] for additional information. A popular and very successful approach to clustering [6, 11, 12] is *spectral clustering*. One first computes an approximation of the first k eigenvectors of the (normalized) Laplacian of G , uses it to embed the vertices of G into k -dimensional Euclidean space \mathbb{R}^k , and then partitions the resulting points via a k -means clustering algorithm. It is an important task for theory to explain the success of spectral clustering. Peng et al. [9] made an important step in this direction recently. They showed that spectral clustering provably works if the $(k+1)$ -th and the k -th eigenvalue of the normalized Laplacian differ sufficiently. In order to explain their result, we need some notation.

The *order k partition constant* $\widehat{\rho}(k)$ of G is defined by

$$\widehat{\rho}(k) \triangleq \min_{\text{partition } (P_1, \dots, P_k) \text{ of } V} \Phi(P_1, \dots, P_k), \quad \text{where} \quad \Phi(Z_1, \dots, Z_k) = \max_{i \in [1:k]} \phi(Z_i).$$

Let $\mathcal{L}_G = I - D^{-1/2} A D^{-1/2}$ be the normalized Laplacian matrix of G , where D is the diagonal degree matrix and A is the adjacency matrix, and let $f_j \in \mathbb{R}^V$ be the eigenvector corresponding to the j -th smallest eigenvalue λ_j of \mathcal{L}_G . The *spectral embedding map* $F : V \rightarrow \mathbb{R}^k$ is defined by

$$F(u) = \frac{1}{\sqrt{d_u}} (f_1(u), \dots, f_k(u))^T, \quad \text{for all vertices } u \in V. \quad (1)$$

Peng et al. [9] construct a k -means instance \mathcal{X}_V by inserting d_u many copies of the vector $F(u)$ into \mathcal{X}_V , for every vertex $u \in V$.

Let \mathcal{X} be a set of vectors of the same dimension. Then

$$\Delta_k(\mathcal{X}) \triangleq \min_{\text{partition } (X_1, \dots, X_k) \text{ of } \mathcal{X}} \sum_{i=1}^k \sum_{x \in X_i} \|x - c_i\|^2, \quad \text{where} \quad c_i = \frac{1}{|X_i|} \sum_{x \in X_i} x,$$

is the optimal cost of clustering \mathcal{X} into k sets. An α -approximate clustering algorithm returns a k -way partition (A_1, \dots, A_k) and centers c_1, \dots, c_k such that

$$\text{Cost}(\{A_i, c_i\}_{i=1}^k) \triangleq \sum_{i=1}^k \sum_{x \in A_i} \|x - c_i\|^2 \leq \alpha \cdot \Delta_k(\mathcal{X}). \quad (2)$$

Theorem 1.1. [9, Theorem 1.2] *Let $k \geq 3$ and (P_1, \dots, P_k) be a k -way partition of V with $\Phi(P_1, \dots, P_k) = \widehat{\rho}(k)$. Let G be a graph that satisfies the gap assumption¹*

$$\Upsilon = \frac{\lambda_{k+1}}{\widehat{\rho}(k)} = 2 \cdot 10^5 \cdot k^3 / \delta, \quad (3)$$

for some $\delta \in (0, 1/2]$. Let (A_1, \dots, A_k) be the k -way partition² of V returned by an α -approximate k -means algorithm applied to \mathcal{X}_V . Then the following statements hold (after suitable renumbering

¹Note that $\lambda_k/2 \leq \widehat{\rho}(k)$, see (6). Thus the assumption implies $\lambda_k/2 \leq \widehat{\rho}(k) \leq \delta \lambda_{k+1} / (2 \cdot 10^5 \cdot k^3)$, i.e., there is a substantial gap between the k -th and the $(k+1)$ -th eigenvalue.

²The k -means algorithm returns a partition of \mathcal{X}_V . One may assume w.l.o.g. that all copies of $F(u)$ are put into the same cluster of \mathcal{X}_V . Thus the algorithm also partitions V .

of one of the partitions):

$$1) \mu(A_i \triangle P_i) \leq \alpha\delta \cdot \mu(P_i) \quad \text{and} \quad 2) \phi(A_i) \leq (1 + 2\alpha\delta) \cdot \phi(P_i) + 2\alpha\delta.$$

Under the stronger gap assumption $\Upsilon = 2 \cdot 10^5 \cdot k^5/\delta$, they showed how to obtain a partition in time $O(m \cdot \text{poly log}(n))$ with essentially the guarantees stated in Theorem 1.1, where $m = |E|$ is the number of edges in G and $n = |V|$ is the number of nodes.

However, their algorithmic result does not analyze the standard spectral clustering paradigm, since it replaces spectral embedding by heat kernel embedding and k -means clustering by locality sensitive hashing. Therefore, their algorithmic result does not explain the success of the standard spectral clustering paradigm.

Our Results: We strengthen the approximation guarantees in Theorem 1.1 by a factor of k and simultaneously weaken the gap assumption. As a consequence, the variant of Lloyd's k -means algorithm analyzed by Ostrovsky et al. [7] applied to³ $\widetilde{\mathcal{X}}_V$ achieves the improved approximation guarantees in time $O(m(k^2 + \frac{\ln n}{\lambda_{k+1}}))$ with constant probability. Table 1 summarizes these results.

Let \mathcal{O} be the set of all k -way partitions (P_1, \dots, P_k) with $\Phi(P_1, \dots, P_k) = \widehat{\rho}(k)$, i.e., the set of all partitions that achieve the order k partition constant. Let

$$\widehat{\rho}_{\text{avr}}(k) \triangleq \min_{(P_1, \dots, P_k) \in \mathcal{O}} \frac{1}{k} \sum_{i=1}^k \phi(P_i)$$

be the *minimal average conductance* over all k -way partitions in \mathcal{O} . Our gap assumption is defined in terms of

$$\Psi \triangleq \frac{\lambda_{k+1}}{\widehat{\rho}_{\text{avr}}(k)}.$$

For the remainder of this paper we denote by (P_1, \dots, P_k) a k -way partition of V that achieves $\widehat{\rho}_{\text{avr}}(k)$. We can now state our main result.

Theorem 1.2 (Main Theorem). *a) (Existence of a Good Clustering) Let G be a graph satisfying*

$$\Psi = 20^4 \cdot k^3/\delta \tag{4}$$

for some $\delta \in (0, 1/2]$ and $k \geq 3$ and let (A_1, \dots, A_k) be the k -way partition output by an α -approximate clustering algorithm applied to the spectral embedding \mathcal{X}_V . Then for every $i \in [1 : k]$ the following two statements hold (after suitable renumbering of one of the partitions):

$$1) \mu(A_i \triangle P_i) \leq \frac{\alpha\delta}{10^3 k} \cdot \mu(P_i) \quad \text{and} \quad 2) \phi(A_i) \leq \left(1 + \frac{2\alpha\delta}{10^3 k}\right) \cdot \phi(P_i) + \frac{2\alpha\delta}{10^3 k}.$$

b) (An Efficient Algorithm) If in addition $k/\delta \geq 10^9$ and⁴ $\Delta_k(\mathcal{X}_V) \geq n^{-O(1)}$, then the variant of Lloyd's algorithm analyzed by Ostrovsky et al. [7] applied to $\widetilde{\mathcal{X}}_V$ returns in time $O(m(k^2 + \frac{\ln n}{\lambda_{k+1}}))$ with constant probability a partition (A_1, \dots, A_k) such that for every $i \in [1 : k]$ the following two statements hold (after suitable renumbering of one of the partitions):

$$3) \mu(A_i \triangle P_i) \leq \frac{2\delta}{10^3 k} \cdot \mu(P_i) \quad \text{and} \quad 4) \phi(A_i) \leq \left(1 + \frac{4\delta}{10^3 k}\right) \cdot \phi(P_i) + \frac{4\delta}{10^3 k}.$$

³ $\widetilde{\mathcal{X}}_V$ is defined as \mathcal{X}_V but in terms of approximate eigenvectors, see Subsection 2.3.

⁴The case $\Delta_k(\mathcal{X}_V) \leq n^{-O(1)}$ constitutes a trivial clustering problem. For technical reasons, we have to exclude too easy inputs.

	Gap Assumption	Partition Quality	Running Time
Peng et al. [9]	$\Upsilon = 2 \cdot 10^5 \cdot k^3 / \delta$	$\mu(A_i \triangle P_i) \leq \alpha \delta \cdot \mu(P_i)$ $\phi(A_i) \leq (1 + 2\alpha \delta) \phi(P_i) + 2\alpha \delta$	Existential result
This paper	$\Psi = 20^4 \cdot k^3 / \delta$	$\mu(A_i \triangle P_i) \leq \frac{\alpha \delta}{10^3 k} \cdot \mu(P_i)$ $\phi(A_i) \leq (1 + \frac{2\alpha \delta}{10^3 k}) \phi(P_i) + \frac{2\alpha \delta}{10^3 k}$	Existential result
Peng et al. [9]	$\Upsilon = 2 \cdot 10^5 \cdot k^5 / \delta$	$\mu(A_i \triangle P_i) \leq \frac{\delta \log^2 k}{k^2} \cdot \mu(P_i)$ $\phi(A_i) \leq \left(1 + \frac{2\delta \log^2 k}{k^2}\right) \phi(P_i) + \frac{2\delta \log^2 k}{k^2}$	$O(m \cdot \text{poly log}(n))$
This paper	$\Psi = 20^4 \cdot k^3 / \delta$ $k/\delta \geq 10^9$ $\Delta_k(\mathcal{X}_V) \geq n^{-O(1)}$	$\mu(A_i \triangle P_i) \leq \frac{2\delta}{10^3 k} \cdot \mu(P_i)$ $\phi(A_i) \leq (1 + \frac{4\delta}{10^3 k}) \phi(P_i) + \frac{4\delta}{10^3 k}$	$O\left(m \left(k^2 + \frac{\ln n}{\lambda_{k+1}}\right)\right)$

Table 1: A comparison of the results in Peng et al. [9] and our results. The parameter $\delta \in (0, 1/2]$ relates the approximation guarantees with the gap assumption.

Part (b) of Theorem 1.2 gives theoretical support for the practical success of spectral clustering based on spectral embedding followed by k -means clustering. Previous papers [5, 9] replaced k -means clustering by other techniques for their algorithmic results.

If $k \leq \text{poly}(\log n)$ and $\lambda_{k+1} \geq \text{poly}(\log n)$, our algorithm works in nearly linear time.

The k -means algorithm in [7] is efficient only for inputs \mathcal{X} for which some partition into k clusters is much better than any partition into $k - 1$ clusters; formally, for inputs \mathcal{X} satisfying $\Delta_k(\mathcal{X}) \leq \varepsilon^2 \cdot \Delta_{k-1}(\mathcal{X})$ for some $\varepsilon \in (0, 6 \cdot 10^{-7}]$. For the proof of part (b) of Theorem 1.2, we show in Section 8 that $\widetilde{\mathcal{X}}_V$ satisfies this assumption.

The *order k conductance constant* $\rho(k)$ is defined by

$$\rho(k) = \min_{\text{disjoint nonempty } Z_1, \dots, Z_k} \Phi(Z_1, \dots, Z_k), \quad \text{where} \quad \Phi(Z_1, \dots, Z_k) = \max_{i \in [1:k]} \phi(Z_i). \quad (5)$$

Lee et al. [5] connected $\rho(k)$ and the k -th smallest eigenvalue of the normalized Laplacian matrix \mathcal{L}_G through the relation

$$\lambda_k/2 \leq \rho(k) \leq O(k^2) \sqrt{\lambda_k}, \quad (6)$$

and in a consecutive work Oveis Gharan and Trevisan [8] showed

$$\rho(k) \leq \widehat{\rho}(k) \leq k\rho(k). \quad (7)$$

In Section 9, we establish an analogous relation for $\widehat{\rho}_{\text{avr}}(k)$. We next introduce our technical contributions and we outline the structure of the paper.

2 Technical Contributions and Structure of the Paper

2.1 Exact Spectral Embedding - Notation

We use the notation adopted by Peng et al. [9]. We refer to the j -th eigenvalue of matrix \mathcal{L}_G by $\lambda_j \triangleq \lambda_j(\mathcal{L}_G)$. The (unit) eigenvector corresponding to λ_j is denoted by $f_j \in \mathbb{R}^V$.

$$\begin{array}{ccc}
\widehat{f}_i = \sum_{j=1}^k \alpha_j^{(i)} f_j & \xrightarrow{\|\widehat{f}_i - \overline{g}_i\|^2 \leq \phi(P_i)/\lambda_{k+1}} & \overline{g}_i = \frac{D^{1/2} \chi_{P_i}}{\sqrt{\mu(P_i)}} = \sum_{j=1}^n \alpha_j^{(i)} f_j \\
\downarrow & & \downarrow \\
f_i = \sum_{j=1}^k \beta_j^{(i)} \widehat{f}_j & \xrightarrow{\|f_i - \widehat{g}_i\|^2 \leq (1 + 3k/\Psi) \cdot k/\Psi} & \widehat{g}_i = \sum_{j=1}^k \beta_j^{(i)} \overline{g}_j
\end{array}$$

Figure 1: The relation between the vectors f_i , \widehat{f}_i , \widehat{g}_i and \overline{g}_i . The vectors $\{f_i\}_{i=1}^n$ are eigenvectors of the normalized Laplacian matrix \mathcal{L}_G of a graph G satisfying $\Psi > 4 \cdot k^{3/2}$. The vectors $\{\overline{g}_i\}_{i=1}^k$ are the normalized characteristic vectors of an optimal partition (P_1, \dots, P_k) . For each $i \in [1 : k]$ the vector \widehat{f}_i is the projection of vector \overline{g}_i onto $\text{span}(f_1, \dots, f_k)$. The vectors \widehat{f}_i and \overline{g}_i are close for $i \in [1 : k]$. It holds $\text{span}(f_1, \dots, f_k) = \text{span}(\widehat{f}_1, \dots, \widehat{f}_k)$ when $\Psi > 4 \cdot k^{3/2}$, and thus we can write $f_i = \sum_{j=1}^k \beta_j^{(i)} \widehat{f}_j$. Moreover, the vectors f_i and $\widehat{g}_i = \sum_{j=1}^k \beta_j^{(i)} \overline{g}_j$ are close for $i \in [1 : k]$.

Let $\overline{g}_i = \frac{D^{1/2} \chi_{P_i}}{\|D^{1/2} \chi_{P_i}\|}$, where χ_{P_i} is the characteristic vector of the subset $P_i \subseteq V$. We note that \overline{g}_i is the normalized characteristic vector of P_i and $\|D^{1/2} \chi_{P_i}\|^2 = \sum_{v \in P_i} \deg(v) = \mu(P_i)$. The Rayleigh quotient is defined by and satisfies

$$\mathcal{R}(\overline{g}_i) \triangleq \frac{\overline{g}_i^T \mathcal{L}_G \overline{g}_i}{\overline{g}_i^T \overline{g}_i} = \frac{1}{\mu(P_i)} \chi_{P_i}^T L \chi_{P_i} = \frac{|E(S, \overline{S})|}{\mu(P_i)} = \phi(P_i),$$

where $L = D - A$ is the graph Laplacian matrix.

The eigenvectors $\{f_i\}_{i=1}^n$ form an orthonormal basis of \mathbb{R}^n . Thus each characteristic vector \overline{g}_i can be expressed as $\overline{g}_i = \sum_{j=1}^n \alpha_j^{(i)} f_j$ for all $i \in [1 : k]$. We define its *projection* onto the first k eigenvectors by $\widehat{f}_i = \sum_{j=1}^k \alpha_j^{(i)} f_j$.

Peng et al. [9] proved that if the gap parameter Υ is large enough then $\text{span}(\{\widehat{f}_i\}_{i=1}^k) = \text{span}(\{\overline{g}_i\}_{i=1}^k)$ and the first k eigenvectors can be expressed by $f_i = \sum_{j=1}^k \beta_j^{(i)} \widehat{f}_j$, for all $i \in [1 : k]$. Moreover, they demonstrated that each vector $\widehat{g}_i = \sum_{j=1}^k \beta_j^{(i)} \overline{g}_j$ approximates the eigenvector f_i , for all $i \in [1 : k]$. We show that similar statements hold with substituted gap parameter Ψ .

We define the estimation centers induced by the spectral embedding by

$$p^{(i)} = \frac{1}{\sqrt{\mu(P_i)}} \left(\beta_i^{(1)}, \dots, \beta_i^{(k)} \right)^T. \quad (8)$$

Our analysis relies on the spectral properties of the following two matrices. Let $\mathbf{F}, \mathbf{B} \in \mathbb{R}^{k \times k}$ be square matrices such that for all indices $i, j \in [1 : k]$ we have

$$\mathbf{F}_{j,i} = \alpha_j^{(i)} \quad \text{and} \quad \mathbf{B}_{j,i} = \beta_j^{(i)}. \quad (9)$$

2.2 Exact Spectral Embedding - Our Results

Our proof of Theorem 1.2 (a) follows the proof-structure of [9, Theorem 1.2] in Peng et al., but improves upon it in essential ways.

Our key technical insight is that the matrices $\mathbf{B}^T \mathbf{B}$ and $\mathbf{B} \mathbf{B}^T$ are close to the identity matrix. We prove this in two steps. In Section 4, we show that the vectors \widehat{g}_i and f_i are close, and then in Section 4 we analyze the column space and row space of matrix \mathbf{B} .

Theorem 2.1 (Matrix $\mathbf{B}\mathbf{B}^\top$ is Close to Identity Matrix). *If $\Psi \geq 10^4 \cdot k^3/\varepsilon^2$ and $\varepsilon \in (0, 1)$ then for all distinct $i, j \in [1 : k]$ it holds*

$$1 - \varepsilon \leq \langle \mathbf{B}_{i,:}, \mathbf{B}_{i,:} \rangle \leq 1 + \varepsilon \quad \text{and} \quad |\langle \mathbf{B}_{i,:}, \mathbf{B}_{j,:} \rangle| \leq \sqrt{\varepsilon}.$$

Theorem 2.1 is key for the improved separation (by a factor of k over [9, Lemma 4.3]) of estimation centers and the bound on the norm of the separation centers.

Lemma 2.2. *If $\Psi = 20^4 \cdot k^3/\delta$ for some $\delta \in (0, 1]$ then for every $i \in [1 : k]$ it holds that*

$$\|p^{(i)}\|^2 \in \left[1 \pm \sqrt{\delta}/4\right] \frac{1}{\mu(P_i)}.$$

Proof. By definition $p^{(i)} = \frac{1}{\sqrt{\mu(P_i)}} \cdot \mathbf{B}_{i,:}$ and Theorem 2.1 yields $\|\mathbf{B}_{i,:}\|^2 \in [1 \pm \sqrt{\delta}/4]$. ■

Lemma 2.3 (Larger Distance Between Estimation Centers). *If $\Psi = 20^4 \cdot k^3/\delta$ for some $\delta \in (0, \frac{1}{2}]$ then for any distinct $i, j \in [1 : k]$ it holds that*

$$\|p^{(i)} - p^{(j)}\|^2 \geq [2 \cdot \min\{\mu(P_i), \mu(P_j)\}]^{-1}.$$

Proof. Since $p^{(i)}$ is a row of matrix B , Theorem 2.1 with $\varepsilon = \sqrt{\delta}/4$ yields

$$\left\langle \frac{p^{(i)}}{\|p^{(i)}\|}, \frac{p^{(j)}}{\|p^{(j)}\|} \right\rangle = \frac{\langle \mathbf{B}_{i,:}, \mathbf{B}_{j,:} \rangle}{\|\mathbf{B}_{i,:}\| \|\mathbf{B}_{j,:}\|} \leq \frac{\sqrt{\varepsilon}}{1 - \varepsilon} = \frac{2\delta^{1/4}}{3}.$$

W.l.o.g. assume that $\|p^{(i)}\|^2 \geq \|p^{(j)}\|^2$, say $\|p^{(j)}\| = \alpha \cdot \|p^{(i)}\|$ for some $\alpha \in (0, 1]$. Then by Lemma 2.2 we have $\|p^{(i)}\|^2 \geq (1 - \sqrt{\delta}/4) \cdot [\min\{\mu(P_i), \mu(P_j)\}]^{-1}$, and hence

$$\begin{aligned} \|p^{(i)} - p^{(j)}\|^2 &= \|p^{(i)}\|^2 + \|p^{(j)}\|^2 - 2 \left\langle \frac{p^{(i)}}{\|p^{(i)}\|}, \frac{p^{(j)}}{\|p^{(j)}\|} \right\rangle \|p^{(i)}\| \|p^{(j)}\| \\ &\geq \left(\alpha^2 - \frac{4\delta^{1/4}}{3} \cdot \alpha + 1 \right) \|p^{(i)}\|^2 \geq [2 \cdot \min\{\mu(P_i), \mu(P_j)\}]^{-1}. \end{aligned}$$
■

The observation that Υ can be replaced by Ψ in all statements in [9] is technically easy. However, this is crucial for the part (b) of Theorem 1.2, since it yields an improved version of [9, Lemma 4.5] showing that a weaker by a factor of k assumption is sufficient. We prove Lemma 2.4 in Section 6.

Lemma 2.4. *Let (P_1, \dots, P_k) and (A_1, \dots, A_k) are partitions of the vector set. Suppose for every permutation $\pi : [1 : k] \rightarrow [1 : k]$ there is an index $i \in [1 : k]$ such that*

$$\mu(A_i \triangle P_{\pi(i)}) \geq \frac{2\varepsilon}{k} \cdot \mu(P_{\pi(i)}), \tag{10}$$

where $\varepsilon \in (0, 1)$ is a parameter. If $\Psi = 20^4 \cdot k^3/\delta$ for some $\delta \in (0, \frac{1}{2}]$, and $\varepsilon \geq 64\alpha \cdot k^3/\Psi$ then

$$\text{Cost}(\{A_i, c_i\}_{i=1}^k) > \frac{2k^2}{\Psi} \alpha.$$

With the above Lemmas in place, the proof of part (a) of Theorem 1.2 is then completed as in [9]. We give more details in Section 3.

Before we turn to part (b) of Theorem 1.2, we consider the variant of Lloyd’s algorithm analyzed by Ostrovsky et al. [7] applied to \mathcal{X}_V . This algorithm is efficient for inputs \mathcal{X} satisfying: some partition into k clusters is much better than any partition into $k - 1$ clusters.

Theorem 2.5. [7, Theorem 4.15] *Assuming that $\Delta_k(\mathcal{X}) \leq \varepsilon^2 \Delta_{k-1}(\mathcal{X})$ for $\varepsilon \in (0, 6 \cdot 10^{-7}]$, there is an algorithm that returns a solution of cost at most $[(1 - \varepsilon^2)/(1 - 37\varepsilon^2)]\Delta_k(\mathcal{X})$ with probability at least $1 - O(\sqrt{\varepsilon})$ in time $O(nkd + k^3d)$.*

In Section 7, we establish the assumption of Ostrovsky et al. [7] for \mathcal{X}_V .

Theorem 2.6 (Normalized Spectral Embedding is ε -separated). *Let G be a graph that satisfies $\Psi = 20^4 \cdot k^3/\delta$, $\delta \in (0, 1/2]$ and $k/\delta \geq 10^9$. Then for $\varepsilon = 6 \cdot 10^{-7}$ it holds*

$$\Delta_k(\mathcal{X}_V) \leq \varepsilon^2 \Delta_{k-1}(\mathcal{X}_V). \quad (11)$$

However, Theorem 2.6 is insufficient for part (b) of Theorem 1.2, since we need a similar result for the set $\widetilde{\mathcal{X}}_V$ formed by approximate eigenvectors. To overcome this issue we build upon the recent work by Boutsidis et al. [2] which shows that running an approximate k -means clustering algorithm on approximate eigenvectors obtained via the power method, yields an additive approximation to solving the k -means clustering problem on exact eigenvectors.

In order to state the connection, we need to introduce some of their notation.

2.3 Approximate Spectral Embedding - Notation

Let $Z \in \mathbb{R}^{n \times k}$ be a matrix whose rows represent n vectors that are to be partitioned into k clusters. For every k -way partition we associate an indicator matrix $X \in \mathbb{R}^{n \times k}$ that satisfies $X_{ij} = 1/\sqrt{|C_j|}$ if the i -th row $Z_{i,:}$ belongs to the j -th cluster C_j , and $X_{ij} = 0$ otherwise. We denote the optimal indicator matrix X_{opt} by

$$X_{\text{opt}} = \arg \min_{X \in \mathbb{R}^{n \times k}} \|Z - XX^T Z\|_F^2 = \arg \min_{X \in \mathbb{R}^{n \times k}} \sum_{j=1}^k \sum_{u \in X_j} \|Z_{u,:} - c_j\|_2^2, \quad (12)$$

where $c_j = (1/|X_j|) \sum_{u \in X_j} Z_{u,:}$ is the center point of cluster C_j .

The normalized Laplacian matrix $\mathcal{L}_G \in \mathbb{R}^{n \times n}$ of a graph G is define by $\mathcal{L}_G = I - \mathcal{A}$, where $\mathcal{A} = D^{-1/2}AD^{-1/2}$ is the normalized adjacency matrix. Let $U_k \in \mathbb{R}^{n \times k}$ be a matrix composed of the first k orthonormal eigenvectors of \mathcal{L}_G corresponding to the smallest eigenvalues $\lambda_1, \dots, \lambda_k$. We define by $Y \triangleq U_k$ the canonical spectral embedding.

Our approximate spectral embedding is computed by the so called “**Power method**”. Let $S \in \mathbb{R}^{n \times k}$ be a matrix whose entries are i.i.d. samples from the standard Gaussian distribution $N(0, 1)$ and p be a positive integer. Then the approximate spectral embedding \widetilde{Y} is defined by the following process:

$$1) \mathcal{B} \triangleq I + \mathcal{A}; \quad 2) \text{ Let } \widetilde{U}\widetilde{\Sigma}\widetilde{V}^T \text{ be the SVD of } \mathcal{B}^p S; \quad \text{and} \quad 3) \widetilde{Y} \triangleq \widetilde{U} \in \mathbb{R}^{n \times k}. \quad (13)$$

We proceed by defining the normalized (approximate) spectral embedding. We construct a matrix $Y' \in \mathbb{R}^{m \times k}$ such that for every vertex $u \in V$ we add $d(u)$ many copies of the normalized

row $U_k(u, :)/\sqrt{d(u)}$ to Y' . Formally, the normalized (approximate) spectral embedding Y' (\widetilde{Y}') is defined by

$$Y' = \begin{pmatrix} \mathbf{1}_{d(1)} \frac{U_k(1,:)}{\sqrt{d(1)}} \\ \dots \\ \mathbf{1}_{d(n)} \frac{U_k(n,:)}{\sqrt{d(n)}} \end{pmatrix}_{m \times k} \quad \text{and} \quad \widetilde{Y}' = \begin{pmatrix} \mathbf{1}_{d(1)} \frac{\widetilde{U}(1,:)}{\sqrt{d(1)}} \\ \dots \\ \mathbf{1}_{d(n)} \frac{\widetilde{U}(n,:)}{\sqrt{d(n)}} \end{pmatrix}_{m \times k}, \quad (14)$$

where $\mathbf{1}_{d(i)}$ is all-one column vector with dimension $d(i)$.

Similarly to (12) we associate to Y' (\widetilde{Y}') an indicator matrix X' (\widetilde{X}') that satisfies $X'_{ij} = 1/\sqrt{\mu(C_j)}$ if the i -th row $Y'_{i,:}$ belongs to the j -th cluster C_j , and $X'_{ij} = 0$ otherwise. We may assume w.l.o.g. that a k -means algorithm outputs an indicator matrix X' such that all copies of row $U_k(v, :)/\sqrt{d(v)}$ belong to the same cluster, for every vertex $v \in V$.

We associate to matrices Y' and \widetilde{Y}' the sets of points \mathcal{X}_V and $\widetilde{\mathcal{X}}_V$ respectively. We present now a key connection between the spectral embedding map $F(\cdot)$, the optimal k -means cost $\Delta_k(\mathcal{X}_V)$ and matrices Y' , X'_{opt} :

$$\left\| Y' - X'_{\text{opt}} (X'_{\text{opt}})^T Y' \right\|_F^2 = \sum_{j=1}^k \sum_{v \in C_j^*} d(v) \|F(v) - c_j^*\|_F^2 = \Delta_k(\mathcal{X}_V), \quad (15)$$

where each center satisfies $c_j^* = \mu(C_j^*)^{-1} \cdot \sum_{v \in C_j^*} d(v) F(v)$ and $F(v) = Y_{v,:}/\sqrt{d(v)}$.

2.4 Approximate Spectral Embedding - Our Results

Our analysis relies on the proof techniques developed in [1, 2]. By adjusting these techniques (c.f. [2, Lemma 5] and [1, Lemma 7]) to our setting, we prove in Subsection 8.1 the following result for the symmetric positive semi-definite matrix \mathcal{B} whose largest k singular values (eigenvalues) correspond to the eigenvectors u_1, \dots, u_k of \mathcal{L}_G .

Lemma 2.7. *Let $\widetilde{U}\widetilde{\Sigma}\widetilde{V}^T$ be the SVD of $\mathcal{B}^p S \in \mathbb{R}^{n \times k}$, where $p \geq 1$ and S is an $n \times k$ matrix of i.i.d. standard Gaussians. Let $\gamma_k = \frac{2-\lambda_{k+1}}{2-\lambda_k} < 1$ and fix $\delta, \epsilon \in (0, 1)$. Then for any $p \geq \ln(8nk/\epsilon\delta)/\ln(1/\gamma_k)$ with probability at least $1 - 2e^{-2n} - 3\delta$ it holds*

$$\left\| U_k U_k^T - \widetilde{U} \widetilde{U}^T \right\|_F \leq \epsilon.$$

We establish several technical Lemmas that combined with Lemma 2.7 allow us to apply the proof techniques in [2, Theorem 6]. More precisely, we prove in Subsection 8.2 that running an approximate k -means algorithm on a normalized approximate spectral embedding \widetilde{Y}' computed by the power method, yields an approximate clustering of the normalized spectral embedding Y' .

Theorem 2.8. *Compute matrix \widetilde{Y}' via the power method with $p \geq \ln(8nk/\epsilon\delta)/\ln(1/\gamma_k)$, where $\gamma_k = (2 - \lambda_{k+1})/(2 - \lambda_k) < 1$. Run on the rows of \widetilde{Y}' an α -approximate k -means algorithm with failure probability δ_α . Let the outcome be a clustering indicator matrix $\widetilde{X}'_\alpha \in \mathbb{R}^{n \times k}$. Then with probability at least $1 - 2e^{-2n} - 3\delta_p - \delta_\alpha$ it holds*

$$\left\| Y' - \widetilde{X}'_\alpha (\widetilde{X}'_\alpha)^T Y' \right\|_F^2 \leq (1 + 4\epsilon) \cdot \alpha \cdot \left\| Y' - X'_{\text{opt}} (X'_{\text{opt}})^T Y' \right\|_F^2 + 4\epsilon^2.$$

Our main technical contribution is to prove, in Subsection 8.3, that $\widetilde{\mathcal{X}}_V$ satisfies the assumption of Ostrovsky et al. [7]. Our analysis builds upon Theorem 2.6 and Theorem 2.8.

Theorem 2.9 (Approximate Normalized Spectral Embedding is ε -separated). *Suppose the gap assumption satisfies $\Psi = 20^4 \cdot k^3/\delta$, $k/\delta \geq 10^9$ for some $\delta \in (0, 1/2]$ and the optimum cost⁵ $\|Y' - X'_{\text{opt}}(X'_{\text{opt}})^T Y'\|_F \geq n^{-O(1)}$. Construct matrix \widetilde{Y}' via the power method with $p \geq \Omega(\frac{\ln n}{\lambda_{k+1}})$. Then for $\varepsilon = 6 \cdot 10^{-7}$ w.h.p it holds $\Delta_k(\widetilde{\mathcal{X}}_V) < 5\varepsilon^2 \cdot \Delta_{k-1}(\widetilde{\mathcal{X}}_V)$.*

Based on the preceding results, we prove part (b) of Theorem 1.2 in Subsection 8.4.

3 The Proof of Part (a) of The Main Theorem:

The proof of part (a.1) builds upon the following Lemmas. Recall that \mathcal{X}_V contains d_u copies of $F(u)$ for each $u \in V$. W.l.o.g. we may restrict attention to clusterings of \mathcal{X}_V that put all copies of $F(u)$ into the same cluster and hence induce a clustering of V . Let (A_1, \dots, A_k) with cluster centers c_1 to c_k be a clustering of V . Its k -means cost is

$$\text{Cost}(\{A_i, c_i\}_{i=1}^k) = \sum_{i=1}^k \sum_{u \in A_i} d_u \|F(u) - c_i\|^2.$$

Lemma 3.1 ((P_1, \dots, P_k) is a good k -means partition). *If $\Psi > 4 \cdot k^{3/2}$ then there are vectors $\{p^{(i)}\}_{i=1}^k$ such that*

$$\text{Cost}(\{P_i, p^{(i)}\}_{i=1}^k) \leq \left(1 + \frac{3k}{\Psi}\right) \cdot \frac{k^2}{\Psi}.$$

Proof. By Theorem 4.1 we have $\|f_i - \widehat{g}_i\|^2 \leq (1 + \frac{3k}{\Psi}) \cdot \frac{k}{\Psi}$ and thus

$$\begin{aligned} \sum_{i=1}^k \sum_{u \in P_i} d_u \|F(u) - c_i^*\|^2 &\leq \sum_{i=1}^k \sum_{u \in P_i} d_u \|F(u) - p^{(i)}\|^2 = \sum_{i=1}^k \sum_{j=1}^k \sum_{u \in P_i} d_u (F(u)_j - p_j^{(i)})^2 \\ &= \sum_{j=1}^k \sum_{i=1}^k \sum_{u \in P_i} (f_j(u) - \widehat{g}_j(u))^2 = \sum_{j=1}^k \|f_j - \widehat{g}_j\|^2 \leq \left(1 + \frac{3k}{\Psi}\right) \cdot \frac{k^2}{\Psi}, \end{aligned}$$

where the k -way partition (P_1, \dots, P_k) achieving $\widehat{\rho}_{\text{avr}}(k)$ has corresponding centers c_1^*, \dots, c_k^* . \blacksquare

Lemma 3.2 (Only partitions close to (P_1, \dots, P_k) are good). *Under the hypothesis of Theorem 1.2, the following holds. If for every permutation $\sigma : [1 : k] \rightarrow [1 : k]$ there exists an index $i \in [1 : k]$ such that*

$$\mu(A_i \triangle P_{\sigma(i)}) \geq \frac{8\alpha\delta}{10^4 k} \cdot \mu(P_{\sigma(i)}). \quad (16)$$

Then it holds that

$$\text{Cost}(\{A_i, c_i\}_{i=1}^k) > \frac{2\alpha k^2}{\Psi}. \quad (17)$$

⁵ $\|Y' - X'_{\text{opt}}(X'_{\text{opt}})^T Y'\|_F \geq n^{-O(1)}$ asserts a multiplicative approximation guarantee in Theorem 2.8.

We note that Lemma 3.2 follows directly by applying Lemma 2.4 with $\varepsilon = 64 \cdot \alpha \cdot k^3 / \Psi$. Substituting these bounds into (2) yields a contradiction, since

$$\frac{2\alpha k^2}{\Psi} < \text{Cost}(\{A_i, c_i\}_{i=1}^k) \leq \alpha \cdot \Delta_k(\mathcal{X}_V) \leq \alpha \cdot \text{Cost}(\{P_i, p^{(i)}\}_{i=1}^k) \leq \left(1 + \frac{3k}{\Psi}\right) \cdot \frac{\alpha k^2}{\Psi}.$$

Therefore, there exists a permutation π (the identity after suitable renumbering of one of the partitions) such that $\mu(A_i \Delta P_i) < \frac{8\alpha\delta}{10^4 k} \cdot \mu(P_i)$ for all $i \in [1 : k]$.

Part (a.2) follows from Part (a.1). Indeed, for $\delta' = 8\delta/10^4$ we have

$$\mu(A_i) \geq \mu(P_i \cap A_i) = \mu(P_i) - \mu(P_i \setminus A_i) \geq \mu(P_i) - \mu(A_i \Delta P_i) \geq \left(1 - \frac{\alpha\delta'}{k}\right) \cdot \mu(P_i)$$

and $|E(A_i, \overline{A_i})| \leq |E(P_i, \overline{P_i})| + \mu(A_i \Delta P_i)$ since every edge that is counted in $|E(A_i, \overline{A_i})|$ but not in $|E(P_i, \overline{P_i})|$ must have an endpoint in $A_i \Delta P_i$. Thus

$$\Phi(A_i) = \frac{|E(A_i, \overline{A_i})|}{\mu(A_i)} \leq \frac{|E(P_i, \overline{P_i})| + \frac{\alpha\delta'}{k} \cdot \mu(P_i)}{(1 - \frac{\alpha\delta'}{k}) \cdot \mu(P_i)} \leq \left(1 + \frac{2\alpha\delta'}{k}\right) \cdot \phi(P_i) + \frac{2\alpha\delta'}{k}.$$

This completes the proof of part (a) of Theorem 1.2.

4 Vectors \widehat{g}_i and f_i are Close

In this section we prove Theorem 4.1. We argue in a similar manner as in [9], but in contrast we use the gap parameter Ψ . For completeness, we show in Subsection 4.1 that the span of the first k eigenvectors of \mathcal{L}_G equals the span of the projections of P_i 's characteristic vectors onto the first k eigenvectors. Then in Subsection 4.2 we express the eigenvectors f_i in terms of \widehat{f}_i and we conclude the proof of Theorem 4.1.

Theorem 4.1. *If $\Psi > 4 \cdot k^{3/2}$ then the vectors $\widehat{g}_i = \sum_{j=1}^k \beta_j^{(i)} \overline{g}_j$, $i \in [1 : k]$, satisfy*

$$\|f_i - \widehat{g}_i\|^2 \leq \left(1 + \frac{3k}{\Psi}\right) \cdot \frac{k}{\Psi}.$$

4.1 Analyzing the Columns of Matrix F

We prove in this subsection the following result that depends on gap parameter Ψ .

Lemma 4.2. *If $\Psi > k^{3/2}$ then the $\text{span}(\{\widehat{f}_i\}_{i=1}^k) = \text{span}(\{f_i\}_{i=1}^k)$ and thus each eigenvector can be expressed as $f_i = \sum_{j=1}^k \beta_j^{(i)} \cdot \widehat{f}_j$ for every $i \in [1 : k]$.*

To prove Lemma 4.2 we build upon the following result shown by Peng et al. [9].

Lemma 4.3. [9, Theorem 1.1 Part 1] *For $P_i \subset V$ let $\overline{g}_i = \frac{D^{1/2} \chi_{P_i}}{\|D^{1/2} \chi_{P_i}\|}$. Then any $i \in [1 : k]$ it holds that*

$$\|\overline{g}_i - \widehat{f}_i\|^2 = \sum_{j=k+1}^n \left(\alpha_j^{(i)}\right)^2 \leq \frac{\mathcal{R}(\overline{g}_i)}{\lambda_{k+1}} = \frac{\phi(P_i)}{\lambda_{k+1}}.$$

Based on the following two results we prove Lemma 4.2.

Lemma 4.4. For every $i \in [1 : k]$ and $p \neq q \in [1 : k]$ it holds that

$$1 - \phi(P_i)/\lambda_{k+1} \leq \|\widehat{f}_i\|^2 = \|\alpha^{(i)}\|^2 \leq 1 \quad \text{and} \quad \left| \langle \widehat{f}_p, \widehat{f}_q \rangle \right| = |\langle \alpha^{(p)}, \alpha^{(q)} \rangle| \leq \frac{\sqrt{\phi(P_p) \cdot \phi(P_q)}}{\lambda_{k+1}}.$$

Proof. The first part follows by Lemma 4.3 and the following chain of inequalities

$$1 - \frac{\phi(P_i)}{\lambda_{k+1}} \leq 1 - \sum_{j=k+1}^n \left(\alpha_j^{(i)} \right)^2 = \|\widehat{f}_i\|^2 = \sum_{j=1}^k \left(\alpha_j^{(i)} \right)^2 \leq \sum_{j=1}^n \left(\alpha_j^{(i)} \right)^2 = 1.$$

We show now the second part. Since $\{f_i\}_{i=1}^n$ are orthonormal eigenvectors we have for all $p \neq q$ that

$$\langle f_p, f_q \rangle = \sum_{l=1}^n \alpha_l^{(p)} \cdot \alpha_l^{(q)} = 0. \quad (18)$$

We combine (18) and Cauchy-Schwarz to obtain

$$\begin{aligned} \left| \langle \widehat{f}_p, \widehat{f}_q \rangle \right| &= \left| \sum_{l=1}^k \alpha_l^{(p)} \cdot \alpha_l^{(q)} \right| = \left| \sum_{l=k+1}^n \alpha_l^{(p)} \cdot \alpha_l^{(q)} \right| \\ &\leq \sqrt{\sum_{l=k+1}^n \left(\alpha_l^{(p)} \right)^2} \cdot \sqrt{\sum_{l=k+1}^n \left(\alpha_l^{(q)} \right)^2} \leq \frac{\sqrt{\phi(P_p) \cdot \phi(P_q)}}{\lambda_{k+1}}. \end{aligned}$$

■

Lemma 4.5. If $\Psi > k^{3/2}$ then the columns $\{\mathbf{F}_{:,i}\}_{i=1}^k$ are linearly independent.

Proof. We show that the columns of matrix \mathbf{F} are almost orthonormal. Consider the symmetric matrix $\mathbf{F}^T \mathbf{F}$. It is known that $\ker(\mathbf{F}^T \mathbf{F}) = \ker(\mathbf{F})$ and that all eigenvalues of matrix $\mathbf{F}^T \mathbf{F}$ are real numbers. We proceed by showing that the smallest eigenvalue $\lambda_{\min}(\mathbf{F}^T \mathbf{F}) > 0$. This would imply that $\ker(\mathbf{F}) = \emptyset$ and hence yields the statement.

By combining Gersgorin Circle Theorem, Lemma 4.4 and Cauchy-Schwarz it holds that

$$\begin{aligned} \lambda_{\min}(\mathbf{F}^T \mathbf{F}) &\geq \min_{i \in [1:k]} \left\{ (\mathbf{F}^T \mathbf{F})_{ii} - \sum_{j \neq i} |(\mathbf{F}^T \mathbf{F})_{ij}| \right\} = \min_{i \in [1:k]} \left\{ \|\alpha^{(i)}\|^2 - \sum_{j \neq i} |\langle \alpha^{(j)}, \alpha^{(i)} \rangle| \right\} \\ &\geq 1 - \sum_{j=1}^k \sqrt{\frac{\phi(P_j)}{\lambda_{k+1}}} \sqrt{\frac{\phi(P_{i^*})}{\lambda_{k+1}}} \geq 1 - \sqrt{k} \sqrt{\sum_{j=1}^k \frac{\phi(P_j)}{\lambda_{k+1}}} \sqrt{\frac{\phi(P_{i^*})}{\lambda_{k+1}}} \geq 1 - \frac{k^{3/2}}{\Psi} > 0, \end{aligned}$$

where $i^* \in [1 : k]$ is the index that minimizes the expression above. ■

We present now the proof of Lemma 4.2.

Proof of Lemma 4.2. Let λ be an arbitrary non-zero vector. Notice that

$$\sum_{i=1}^k \lambda_i \cdot \widehat{f}_i = \sum_{i=1}^k \lambda_i \sum_{j=1}^k \alpha_j^{(i)} f_j = \sum_{j=1}^k \left(\sum_{i=1}^k \lambda_i \alpha_j^{(i)} \right) f_j = \sum_{j=1}^k \gamma_j f_j, \quad \text{where } \gamma_j = \langle \mathbf{F}_{j,:}, \lambda \rangle. \quad (19)$$

By Lemma 4.5 the columns $\{\mathbf{F}_{:,i}\}_{i=1}^k$ are linearly independent and since $\gamma = \mathbf{F}\lambda$, it follows at least one component $\gamma_j \neq 0$. Therefore the vectors $\{\widehat{f}_i\}_{i=1}^k$ are linearly independent and span \mathbb{R}^k . ■

4.2 Analyzing Eigenvectors f in terms of \widehat{f}_j

To prove Theorem 4.1 we establish next the following result.

Lemma 4.6. *If $\Psi > k^{3/2}$ then for $i \in [k]$ it holds*

$$\left(1 + \frac{2k}{\Psi}\right)^{-1} \leq \sum_{j=1}^k \left(\beta_j^{(i)}\right)^2 \leq \left(1 - \frac{2k}{\Psi}\right)^{-1}.$$

Proof. We show now the upper bound. By Lemma 4.2 $f_i = \sum_{j=1}^k \beta_j^{(i)} \widehat{f}_j$ for all $i \in [1 : k]$ and thus

$$\begin{aligned} 1 &= \|f_i\|^2 = \left\langle \sum_{a=1}^k \beta_a^{(i)} \widehat{f}_a, \sum_{b=1}^k \beta_b^{(i)} \widehat{f}_b \right\rangle \\ &= \sum_{j=1}^k \left(\beta_j^{(i)}\right)^2 \|\widehat{f}_j\|^2 + \sum_{a=1}^k \sum_{b \neq a}^k \beta_a^{(i)} \beta_b^{(i)} \langle \widehat{f}_a, \widehat{f}_b \rangle \\ &\stackrel{(\star)}{\geq} \left(1 - \frac{2k}{\Psi}\right) \cdot \sum_{j=1}^k \left(\beta_j^{(i)}\right)^2. \end{aligned}$$

To prove the inequality (\star) we consider the two terms separately.

By Lemma 4.4, $\|\widehat{f}_j\|^2 \geq 1 - \phi(P_j)/\lambda_{k+1}$. We then apply $\sum_i a_i b_i \leq (\sum_i a_i)(\sum_i b_i)$ for all non-negative vectors a, b and obtain

$$\sum_{j=1}^k \left(\beta_j^{(i)}\right)^2 \left(1 - \frac{\phi(P_j)}{\lambda_{k+1}}\right) = \sum_{j=1}^k \left(\beta_j^{(i)}\right)^2 - \sum_{j=1}^k \left(\beta_j^{(i)}\right)^2 \frac{\phi(P_j)}{\lambda_{k+1}} \geq \left(1 - \frac{k}{\Psi}\right) \sum_{j=1}^k \left(\beta_j^{(i)}\right)^2.$$

Again by Lemma 4.4, we have $|\langle \widehat{f}_a, \widehat{f}_b \rangle| \leq \sqrt{\phi(P_a)\phi(P_b)}/\lambda_{k+1}$, and by Cauchy-Schwarz it holds

$$\begin{aligned} \sum_{a=1}^k \sum_{b \neq a}^k \beta_a^{(i)} \beta_b^{(i)} \langle \widehat{f}_a, \widehat{f}_b \rangle &\geq - \sum_{a=1}^k \sum_{b \neq a}^k |\beta_a^{(i)}| \cdot |\beta_b^{(i)}| \cdot |\langle \widehat{f}_a, \widehat{f}_b \rangle| \\ &\geq - \frac{1}{\lambda_{k+1}} \sum_{a=1}^k \sum_{b \neq a}^k |\beta_a^{(i)}| \sqrt{\phi(P_a)} \cdot |\beta_b^{(i)}| \sqrt{\phi(P_b)} \\ &\geq - \frac{1}{\lambda_{k+1}} \left(\sum_{j=1}^k |\beta_j^{(i)}| \sqrt{\phi(P_j)} \right)^2 \geq - \frac{k}{\Psi} \cdot \sum_{j=1}^k \left(\beta_j^{(i)}\right)^2. \end{aligned}$$

The lower bound follows by analogous arguments. ■

We are ready now to prove Theorem 4.1.

Proof of Theorem 4.1. By Lemma 4.2, we have $f_i = \sum_{j=1}^k \beta_j^{(i)} \widehat{f}_j$ and recall that $\widehat{g}_i = \sum_{j=1}^k \beta_j^{(i)} \overline{g_j}$ for all $i \in [1 : k]$. We combine triangle inequality, Cauchy-Schwarz, Lemma 4.3 and Lemma 4.6 to

obtain

$$\begin{aligned}
\|f_i - \hat{g}_i\|^2 &= \left\| \sum_{j=1}^k \beta_j^{(i)} (\hat{f}_j - \bar{g}_j) \right\|^2 \leq \left(\sum_{j=1}^k |\beta_j^{(i)}| \cdot \|\hat{f}_j - \bar{g}_j\| \right)^2 \\
&\leq \left(\sum_{j=1}^k (\beta_j^{(i)})^2 \right) \cdot \left(\sum_{j=1}^k \|\hat{f}_j - \bar{g}_j\|^2 \right) \leq \left(1 - \frac{2k}{\Psi} \right)^{-1} \left(\frac{1}{\lambda_{k+1}} \sum_{j=1}^k \phi(P_j) \right) \\
&= \left(1 - \frac{2k}{\Psi} \right)^{-1} \cdot \frac{k}{\Psi} \leq \left(1 + \frac{3k}{\Psi} \right) \cdot \frac{k}{\Psi},
\end{aligned}$$

where the last inequality uses $\Psi > 4 \cdot k$. ■

5 Spectral Properties of Matrix \mathbf{B}

In this section, we present the proof of Theorem 2.1. In Subsection 5.1, we analyze the column space of \mathbf{B} and we prove in Lemma 5.3 that $1 - \varepsilon \leq \langle \mathbf{B}_{i,:}, \mathbf{B}_{i,:} \rangle \leq 1 + \varepsilon$ for all i . Then in Subsection 5.2, we analyze the row space of \mathbf{B} and we derive Lemma 5.4 that shows $|\langle \mathbf{B}_{i,:}, \mathbf{B}_{j,:} \rangle| \leq \sqrt{\varepsilon}$ for all i, j .

5.1 Analyzing the Column Space of Matrix \mathbf{B}

We show below that the matrix $\mathbf{B}^T \mathbf{B}$ is close to the identity matrix.

Lemma 5.1. (*Columns*) If $\Psi > 4 \cdot k^{3/2}$ then for all distinct $i, j \in [1 : k]$ it holds

$$1 - \frac{3k}{\Psi} \leq \langle \mathbf{B}_{:,i}, \mathbf{B}_{:,i} \rangle \leq 1 + \frac{3k}{\Psi} \quad \text{and} \quad |\langle \mathbf{B}_{:,i}, \mathbf{B}_{:,j} \rangle| \leq 4\sqrt{\frac{k}{\Psi}}.$$

Proof. By Lemma 4.6 it holds that

$$1 - \frac{3k}{\Psi} \leq \langle \mathbf{B}_{:,i}, \mathbf{B}_{:,i} \rangle = \sum_{j=1}^k (\beta_j^{(i)})^2 \leq 1 + \frac{3k}{\Psi}.$$

Recall that $\hat{g}_i = \sum_{j=1}^k \beta_j^{(i)} \cdot \bar{g}_j$. Moreover, since the eigenvectors $\{f_i\}_{i=1}^k$ and the characteristic vectors $\{\bar{g}_i\}_{i=1}^k$ are orthonormal by combining Cauchy-Schwarz and by Theorem 4.1 it holds

$$\begin{aligned}
|\langle \mathbf{B}_{:,i}, \mathbf{B}_{:,j} \rangle| &= \sum_{l=1}^k \beta_l^{(i)} \beta_l^{(j)} = \left\langle \sum_{a=1}^k \beta_a^{(i)} \cdot \bar{g}_a, \sum_{b=1}^k \beta_b^{(j)} \cdot \bar{g}_b \right\rangle = \langle \hat{g}_i, \hat{g}_j \rangle \\
&= \langle (\hat{g}_i - f_i) + f_i, (\hat{g}_j - f_j) + f_j \rangle \\
&= \langle \hat{g}_i - f_i, \hat{g}_j - f_j \rangle + \langle \hat{g}_i - f_i, f_j \rangle + \langle f_i, \hat{g}_j - f_j \rangle \\
&\leq \|\hat{g}_i - f_i\| \cdot \|\hat{g}_j - f_j\| + \|\hat{g}_i - f_i\| + \|\hat{g}_j - f_j\| \\
&\leq \left(1 + \frac{3k}{\Psi} \right) \cdot \frac{k}{\Psi} + 2\sqrt{\left(1 + \frac{3k}{\Psi} \right) \cdot \frac{k}{\Psi}} \leq 4\sqrt{\frac{k}{\Psi}}.
\end{aligned}$$
■

Using a stronger gap assumption we show that the columns of matrix \mathbf{B} are linearly independent.

Lemma 5.2. *If $\Psi > 25 \cdot k^3$ then the columns $\{\mathbf{B}_{:,i}\}_{i=1}^k$ are linearly independent.*

Proof. Since $\ker(\mathbf{B}) = \ker(\mathbf{B}^T \mathbf{B})$ and $\mathbf{B}^T \mathbf{B}$ is SPSD⁶ matrix, it suffices to show that the smallest eigenvalue

$$\lambda(\mathbf{B}^T \mathbf{B}) = \min_{x \neq 0} \frac{x^T \mathbf{B}^T \mathbf{B} x}{x^T x} > 0.$$

By Lemma 5.1,

$$\sum_{i=1}^k \sum_{j \neq i}^k |x_i| |x_j| \left| \langle \beta^{(i)}, \beta^{(j)} \rangle \right| \leq 4 \sqrt{\frac{k}{\Psi}} \left(\sum_{i=1}^k |x_i| \right)^2 \leq \|x\|^2 \cdot 4k \sqrt{\frac{k}{\Psi}},$$

and

$$\begin{aligned} x^T \mathbf{B}^T \mathbf{B} x &= \left\langle \sum_{i=1}^k x_i \beta^{(i)}, \sum_{j=1}^k x_j \beta^{(j)} \right\rangle = \sum_{i=1}^k x_i^2 \|\beta^{(i)}\|^2 + \sum_{i=1}^k \sum_{j \neq i}^k x_i x_j \langle \beta^{(i)}, \beta^{(j)} \rangle \\ &\geq \left(1 - \frac{3k}{\Psi}\right) \|x\|^2 - \sum_{i=1}^k \sum_{j \neq i}^k |x_i| |x_j| \left| \langle \beta^{(i)}, \beta^{(j)} \rangle \right| \geq \left(1 - 5k \sqrt{\frac{k}{\Psi}}\right) \cdot \|x\|^2. \end{aligned}$$

Therefore $\lambda(\mathbf{B}^T \mathbf{B}) > 0$ and the statement follows. ■

5.2 Analyzing the Row Space of Matrix \mathbf{B}

In this subsection we show that the matrix $\mathbf{B} \mathbf{B}^T$ is close to the identity matrix. We bound now the squared L_2 norm of the rows in matrix \mathbf{B} , i.e. the diagonal entries in matrix $\mathbf{B} \mathbf{B}^T$.

Lemma 5.3. (Rows) *If $\Psi \geq 400 \cdot k^3 / \varepsilon^2$ and $\varepsilon \in (0, 1)$ then for all distinct $i, j \in [1 : k]$ it holds*

$$1 - \varepsilon \leq \langle \mathbf{B}_{i,:}, \mathbf{B}_{i,:} \rangle \leq 1 + \varepsilon.$$

Proof. We show that the eigenvalues of matrix $\mathbf{B} \mathbf{B}^T$ are concentrated around 1. This would imply that $\chi_i^T \mathbf{B} \mathbf{B}^T \chi_i = \langle \mathbf{B}_{i,:}, \mathbf{B}_{i,:} \rangle \approx 1$, where χ_i is a characteristic vector. By Lemma 5.1 we have

$$\left(1 - \frac{3k}{\Psi}\right)^2 \leq \left(\beta^{(i)}\right)^T \cdot \mathbf{B} \mathbf{B}^T \cdot \beta^{(i)} = \|\beta^{(i)}\|^4 + \sum_{j \neq i}^k \langle \beta^{(j)}, \beta^{(i)} \rangle^2 \leq \left(1 + \frac{3k}{\Psi}\right)^2 + \frac{16k^2}{\Psi} \leq 1 + \frac{23k^2}{\Psi}$$

and

$$\left| \left(\beta^{(i)}\right)^T \cdot \mathbf{B} \mathbf{B}^T \cdot \beta^{(j)} \right| \leq \sum_{l=1}^k \left| \langle \beta^{(i)}, \beta^{(l)} \rangle \right| \left| \langle \beta^{(l)}, \beta^{(j)} \rangle \right| \leq 8 \left(1 + \frac{3k}{\Psi}\right) \sqrt{\frac{k}{\Psi}} + 16 \frac{k^2}{\Psi} \leq 11 \sqrt{\frac{k}{\Psi}}.$$

By Lemma 5.2 every vector $x \in \mathbb{R}^k$ can be expressed as $x = \sum_{i=1}^k \gamma_i \beta^{(i)}$.

$$\begin{aligned} x^T \mathbf{B} \mathbf{B}^T x &= \sum_{i=1}^k \gamma_i \left(\beta^{(i)}\right)^T \cdot \mathbf{B} \mathbf{B}^T \cdot \sum_{j=1}^k \gamma_j \beta^{(j)} \\ &= \sum_{i=1}^k \gamma_i^2 \left(\beta^{(i)}\right)^T \cdot \mathbf{B} \mathbf{B}^T \cdot \beta^{(i)} + \sum_{i=1}^k \sum_{j \neq i}^k \gamma_i \gamma_j \left(\beta^{(i)}\right)^T \cdot \mathbf{B} \mathbf{B}^T \cdot \beta^{(j)} \\ &\geq \left(1 - \frac{23k^2}{\Psi} - 11k \sqrt{\frac{k}{\Psi}}\right) \|\gamma\|^2 \geq \left(1 - 14k \sqrt{\frac{k}{\Psi}}\right) \|\gamma\|^2. \end{aligned}$$

⁶We denote by SPSD the class of symmetric positive semi-definite matrices.

and

$$x^T x = \sum_{i=1}^k \sum_{j=1}^k \gamma_i \gamma_j \langle \beta^{(i)}, \beta^{(j)} \rangle = \sum_{i=1}^k \gamma_i^2 \|\beta^{(i)}\|^2 + \sum_{i=1}^k \sum_{j \neq i}^k \gamma_i \gamma_j \langle \beta^{(i)}, \beta^{(j)} \rangle$$

By Lemma 5.1 we have $\left| \sum_{i=1}^k \sum_{j \neq i}^k \gamma_i \gamma_j \langle \beta^{(i)}, \beta^{(j)} \rangle \right| \leq \|\gamma\|^2 \cdot 4k \sqrt{\frac{k}{\Psi}}$ and $\|\beta^{(i)}\|^2 \leq 1 + \frac{3k}{\Psi}$. Thus it holds

$$\left(1 - 5k \sqrt{\frac{k}{\Psi}}\right) \|\gamma\|^2 \leq x^T x \leq \left(1 + 5k \sqrt{\frac{k}{\Psi}}\right) \|\gamma\|^2.$$

Therefore

$$1 - 20k \sqrt{\frac{k}{\Psi}} \leq \lambda(\mathbf{B}\mathbf{B}^T) \leq 1 + 20k \sqrt{\frac{k}{\Psi}}.$$

■

We have now established the first part of Theorem 2.1. We turn to the second part and restate it in the following Lemma.

Lemma 5.4. (Rows) If $\Psi \geq 10^4 \cdot k^3 / \varepsilon^2$ and $\varepsilon \in (0, 1)$ then for all distinct $i, j \in [1 : k]$ it holds

$$|\langle \mathbf{B}_{i,:}, \mathbf{B}_{j,:} \rangle| \leq \sqrt{\varepsilon}.$$

To prove Lemma 5.4 we establish the following three Lemmas. Before stating them we need some notation that is inspired by Lemma 5.1.

Definition 5.5. Let $\mathbf{B}^T \mathbf{B} = \mathbf{I} + \mathbf{E}$, where $|\mathbf{E}_{ij}| \leq 4\sqrt{\frac{k}{\Psi}}$ and \mathbf{E} is symmetric matrix. Then we have

$$(\mathbf{B}\mathbf{B}^T)^2 = \mathbf{B}(\mathbf{I} + \mathbf{E})\mathbf{B}^T = \mathbf{B}\mathbf{B}^T + \mathbf{B}\mathbf{E}\mathbf{B}^T.$$

Lemma 5.6. If $\Psi \geq 40^2 \cdot k^3 / \varepsilon^2$ and $\varepsilon \in (0, 1)$ then all eigenvalues of matrix $\mathbf{B}\mathbf{E}\mathbf{B}^T$ satisfy

$$|\lambda(\mathbf{B}\mathbf{E}\mathbf{B}^T)| \leq \varepsilon/5.$$

Proof. Let $z = \mathbf{B}^T x$. We upper bound the quadratic form

$$|x^T \mathbf{B}\mathbf{E}\mathbf{B}^T x| = |z^T \mathbf{E} z| \leq \sum_{ij} |\mathbf{E}_{ij}| |z_i| |z_j| \leq 4\sqrt{\frac{k}{\Psi}} \cdot \left(\sum_{i=1}^k |z_i| \right)^2 \leq \|z\|^2 \cdot 4k \sqrt{\frac{k}{\Psi}}.$$

By Lemma 5.3 we have $1 - \varepsilon \leq \lambda(\mathbf{B}\mathbf{B}^T) \leq 1 + \varepsilon$ and since $\|z\|^2 = \frac{x \mathbf{B}\mathbf{B}^T x}{x^T x} \cdot \|x\|^2$ it follows that

$$\frac{\|z\|^2}{1 + \varepsilon} \leq \|x\|^2 \leq \frac{\|z\|^2}{1 - \varepsilon},$$

and hence

$$|\lambda(\mathbf{B}\mathbf{E}\mathbf{B}^T)| \leq \max_x \frac{|x^T \mathbf{B}\mathbf{E}\mathbf{B}^T x|}{x^T x} \leq 4(1 + \varepsilon) \cdot k \sqrt{\frac{k}{\Psi}} \leq \varepsilon/5.$$

■

Lemma 5.7. Suppose $\{u_i\}_{i=1}^k$ is orthonormal basis and the square matrix \mathbf{U} has u_i as its i -th column. Then $\mathbf{U}^T \mathbf{U} = \mathbf{I} = \mathbf{U}\mathbf{U}^T$.

Proof. Notice that by the definition of \mathbf{U} it holds $\mathbf{U}^T \mathbf{U} = \mathbf{I}$. Moreover, the matrix \mathbf{U}^{-1} exists and thus $\mathbf{U}^T = \mathbf{U}^{-1}$. Therefore, we have $\mathbf{U} \mathbf{U}^T = \mathbf{I}$ as claimed. \blacksquare

Lemma 5.8. *If $\Psi \geq 40^2 \cdot k^3 / \varepsilon^2$ and $\varepsilon \in (0, 1)$ then it holds $|(\mathbf{BEB}^T)_{ij}| \leq \varepsilon/5$ for every $i, j \in [1 : k]$.*

Proof. Notice that \mathbf{BEB}^T is symmetric matrix, since E is symmetric. By SVD Theorem there is an orthonormal basis $\{u_i\}_{i=1}^k$ such that $\mathbf{BEB}^T = \sum_{i=1}^k \lambda_i(\mathbf{BEB}^T) \cdot u_i u_i^T$. Thus, it suffices to bound the expression

$$|(\mathbf{BEB}^T)_{ij}| \leq \sum_{l=1}^k |\lambda_l(\mathbf{BEB}^T)| \cdot |(u_l u_l^T)_{ij}|.$$

By Lemma 5.7 we have

$$\sum_{l=1}^k |(u_l)_i| \cdot |(u_l)_j| \leq \sqrt{\|\mathbf{U}_{i,:}\|^2} \sqrt{\|\mathbf{U}_{j,:}\|^2} = 1.$$

We apply now Lemma 5.6 to obtain

$$\sum_{l=1}^k |\lambda_l(\mathbf{BEB}^T)| \cdot |(u_l u_l^T)_{ij}| \leq \frac{\varepsilon}{5} \cdot \sum_{l=1}^k |(u_l)_i| \cdot |(u_l)_j| \leq \frac{\varepsilon}{5}.$$

\blacksquare

We are ready now to prove Lemma 5.4, i.e. $|\langle \mathbf{B}_{i,:}, \mathbf{B}_{j,:} \rangle| \leq \sqrt{\varepsilon}$ for all $i \neq j$.

Proof of Lemma 5.4. By Definition 5.5 we have $(\mathbf{BB}^T)^2 = \mathbf{BB}^T + \mathbf{BEB}^T$. Observe that the (i, j) -th entry of matrix \mathbf{BB}^T is equal to the inner product between the i -th and j -th row of matrix \mathbf{B} , i.e. $(\mathbf{BB}^T)_{ij} = \langle \mathbf{B}_{i,:}, \mathbf{B}_{j,:} \rangle$. Moreover, we have

$$\left[(\mathbf{BB}^T)^2 \right]_{ij} = \sum_{l=1}^k (\mathbf{BB}^T)_{i,l} (\mathbf{BB}^T)_{l,j} = \sum_{l=1}^k \langle \mathbf{B}_{i,:}, \mathbf{B}_{l,:} \rangle \langle \mathbf{B}_{l,:}, \mathbf{B}_{j,:} \rangle.$$

For the entries on the main diagonal, it holds

$$\langle \mathbf{B}_{i,:}, \mathbf{B}_{i,:} \rangle^2 + \sum_{l \neq i} \langle \mathbf{B}_{i,:}, \mathbf{B}_{l,:} \rangle^2 = [(\mathbf{BB}^T)^2]_{ii} = [\mathbf{BB}^T + \mathbf{BEB}^T]_{ii} = \langle \mathbf{B}_{i,:}, \mathbf{B}_{i,:} \rangle + (\mathbf{BEB}^T)_{ii},$$

and hence by applying Lemma 5.3 with $\varepsilon' = \varepsilon/5$ and Lemma 5.8 with $\varepsilon' = \varepsilon$ we obtain

$$\langle \mathbf{B}_{i,:}, \mathbf{B}_{j,:} \rangle^2 \leq \sum_{l \neq i} \langle \mathbf{B}_{i,:}, \mathbf{B}_{l,:} \rangle^2 \leq \left(1 + \frac{\varepsilon}{5}\right) + \frac{\varepsilon}{5} - \left(1 - \frac{\varepsilon}{5}\right)^2 \leq \varepsilon.$$

\blacksquare

6 Proof of Lemma 2.4

In this section, we prove Lemma 2.4. Our key technical contribution gives in Lemma 6.3 an improved lower bound on the k -means cost in the setting of Lemma 6.2. This yields an improved version of [9, Lemma 4.5] showing that a weaker by a factor of k assumption suffices (c.f. Lemma 2.4).

Our analysis combines the proof techniques developed in [9] with our strengthened results that depend on the gap parameter Ψ . We start by establishing a Corollary of Lemma 2.3.

Corollary 6.1. *Let $\Psi = 20^4 \cdot k^3/\delta$ for some $\delta \in (0, 1/2]$. Suppose c_i is the center of a cluster A_i . If $\|c_i - p^{(i_1)}\| \geq \|c_i - p^{(i_2)}\|$ then it holds*

$$\|c_i - p^{(i_1)}\|^2 \geq \frac{1}{4} \|p^{(i_1)} - p^{(i_2)}\|^2 \geq [8 \cdot \min\{\mu(P_{i_1}), \mu(P_{i_2})\}]^{-1}.$$

We restate now [9, Lemma B.2] whose analysis crucially relies on a function σ defined by

$$\sigma(l) = \arg \max_{j \in [1:k]} \frac{\mu(A_l \cap P_j)}{\mu(P_j)}. \quad (20)$$

Lemma 6.2. [9, Lemma B.2] *Let (P_1, \dots, P_k) and (A_1, \dots, A_k) be partitions of the vector set. Suppose for every permutation $\pi : [1:k] \rightarrow [1:k]$ there is an index $i \in [1:k]$ such that*

$$\mu(A_i \triangle P_{\pi(i)}) \geq 2\varepsilon \cdot \mu(P_{\pi(i)}), \quad (21)$$

where $\varepsilon \in (0, 1/2)$ is a parameter. Then one of the following three statements holds:

1. *If σ is a permutation and $\mu(P_{\sigma(i)} \setminus A_i) \geq \varepsilon \cdot \mu(P_{\sigma(i)})$, then for every index $j \neq i$ there is a real $\varepsilon_j \geq 0$ such that*

$$\mu(A_j \cap P_{\sigma(j)}) \geq \mu(A_j \cap P_{\sigma(i)}) \geq \varepsilon_j \cdot \min\{\mu(P_{\sigma(j)}), \mu(P_{\sigma(i)})\},$$

and $\sum_{j \neq i} \varepsilon_j \geq \varepsilon$.

2. *If σ is a permutation and $\mu(A_i \setminus P_{\sigma(i)}) \geq \varepsilon \cdot \mu(P_{\sigma(i)})$, then for every $j \neq i$ there is a real $\varepsilon_j \geq 0$ such that*

$$\mu(A_i \cap P_{\sigma(i)}) \geq \varepsilon_j \cdot \mu(P_{\sigma(i)}), \quad \mu(A_i \cap P_{\sigma(j)}) \geq \varepsilon_j \cdot \mu(P_{\sigma(i)}),$$

and $\sum_{j \neq i} \varepsilon_j \geq \varepsilon$.

3. *If σ is not a permutation, then there is an index $\ell \notin \{\sigma(1), \dots, \sigma(k)\}$ and for every index j there is a real $\varepsilon_j \geq 0$ such that*

$$\mu(A_j \cap P_{\sigma(j)}) \geq \mu(A_j \cap P_\ell) \geq \varepsilon_j \cdot \min\{\mu(P_{\sigma(j)}), \mu(P_\ell)\},$$

and $\sum_{j=1}^k \varepsilon_j = 1$.

We prove below the improved lower bound on the k -means cost.

Lemma 6.3. *Suppose the hypothesis of Lemma 6.2 is satisfied and $\Psi = 20^4 \cdot k^3/\delta$ for some $\delta \in (0, 1/2]$. Then it holds*

$$\text{Cost}(\{A_i, c_i\}_{i=1}^k) \geq \frac{\varepsilon}{16} - \frac{2k^2}{\Psi}.$$

Proof. By definition

$$\text{Cost}(\{A_i, c_i\}_{i=1}^k) = \sum_{i=1}^k \sum_{j=1}^k \sum_{u \in A_i \cap P_j} d_u \|F(u) - c_i\|^2 \triangleq \Lambda. \quad (22)$$

Since for every vectors $x, y, z \in \mathbb{R}^k$ it holds

$$2 \left(\|x - y\|^2 + \|z - y\|^2 \right) \geq (\|x - y\| + \|z - y\|)^2 \geq \|x - z\|^2,$$

we have for all indices $i, j \in [1 : k]$ that

$$\|F(u) - c_i\|^2 \geq \frac{\|p^{(j)} - c_i\|^2}{2} - \|F(u) - p^{(j)}\|^2. \quad (23)$$

Our proof proceeds by considering three cases. Let $i \in [1 : k]$ be the index from the hypothesis in Lemma 6.2.

Case 1. Suppose the first conclusion of Lemma 6.2 holds. For every index $j \neq i$ let

$$p^{\gamma(j)} = \begin{cases} p^{\sigma(j)} & , \text{ if } \|p^{\sigma(j)} - c_j\| \geq \|p^{\sigma(i)} - c_j\|; \\ p^{\sigma(i)} & , \text{ otherwise.} \end{cases}$$

Then by combining (23), Corollary 6.1 and Lemma 3.1, we have

$$\begin{aligned} \Lambda &\geq \frac{1}{2} \sum_{j \neq i} \sum_{u \in A_j \cap P_{\gamma(j)}} d_u \|p^{\gamma(j)} - c_j\|^2 - \sum_{j \neq i} \sum_{u \in A_j \cap P_{\gamma(j)}} \|F(u) - p^{\gamma(j)}\|^2 \\ &\geq \frac{1}{16} \sum_{j \neq i} \frac{\mu(A_j \cap P_{\gamma(j)})}{\min\{\mu(P_{\sigma(i)}), \mu(P_{\sigma(j)})\}} - \left(1 + \frac{3k}{\Psi}\right) \cdot \frac{k^2}{\Psi} \geq \frac{\varepsilon}{16} - \frac{2k^2}{\Psi}. \end{aligned}$$

Case 2. Suppose the second conclusion of Lemma 6.2 holds. Notice that if $\mu(A_i \cap P_{\sigma(i)}) \leq (1 - \varepsilon) \cdot \mu(P_{\sigma(i)})$ then $\mu(P_{\sigma(i)} \setminus A_i) \geq \varepsilon \cdot \mu(P_{\sigma(i)})$ and thus we can argue as in Case 1. Hence, we can assume that it holds

$$\mu(A_i \cap P_{\sigma(i)}) \geq (1 - \varepsilon) \cdot \mu(P_{\sigma(i)}). \quad (24)$$

We proceed by analyzing two subcases.

a) If $\|p^{\sigma(j)} - c_i\| \geq \|p^{\sigma(i)} - c_i\|$ holds for all $j \neq i$ then by combining (23), Corollary 6.1 and Lemma 3.1 it follows

$$\begin{aligned} \Lambda &\geq \frac{1}{2} \sum_{j \neq i} \sum_{u \in A_i \cap P_{\sigma(j)}} d_u \|p^{\sigma(j)} - c_i\|^2 - \sum_{j \neq i} \sum_{u \in A_i \cap P_{\sigma(j)}} \|F(u) - p^{\sigma(j)}\|^2 \\ &\geq \frac{1}{2} \sum_{j \neq i} \frac{\mu(A_i \cap P_{\sigma(j)})}{\min\{\mu(P_{\sigma(i)}), \mu(P_{\sigma(j)})\}} - \left(1 + \frac{3k}{\Psi}\right) \cdot \frac{k^2}{\Psi} \geq \frac{\varepsilon}{16} - \frac{2k^2}{\Psi}. \end{aligned}$$

b) Suppose there is an index $j \neq i$ such that $\|p^{\sigma(j)} - c_i\| < \|p^{\sigma(i)} - c_i\|$. Then by triangle inequality combined with Corollary 6.1 we have

$$\|p^{\sigma(i)} - c_i\|^2 \geq \frac{1}{4} \|p^{\sigma(i)} - p^{\sigma(j)}\|^2 \geq [8 \cdot \min\{\mu(P_{\sigma(i)}), \mu(P_{\sigma(j)})\}]^{-1}.$$

Thus, by combining (23), (24) and Lemma 3.1 we obtain

$$\begin{aligned}\Lambda &\geq \frac{1}{2} \sum_{u \in A_i \cap P_{\sigma(i)}} d_u \|p^{\sigma(i)} - c_i\|^2 - \sum_{u \in A_i \cap P_{\sigma(i)}} d_u \|F(u) - p^{\sigma(i)}\|^2 \\ &\geq \frac{1}{16} \cdot \frac{\mu(A_i \cap P_{\sigma(i)})}{\min\{\mu(P_{\sigma(i)}), \mu(P_{\sigma(j)})\}} - \left(1 + \frac{3k}{\Psi}\right) \cdot \frac{k^2}{\Psi} \geq \frac{1-\varepsilon}{16} - \frac{2k^2}{\Psi}.\end{aligned}$$

Case 3. Suppose the third conclusion of Lemma 6.2 holds, i.e., σ is not a permutation. Then there is an index $\ell \in [1 : k] \setminus \{\sigma(1), \dots, \sigma(k)\}$ and for every index $j \in [1 : k]$ let

$$p^{\gamma(j)} = \begin{cases} p^\ell & , \text{ if } \|p^\ell - c_j\| \geq \|p^{\sigma(j)} - c_j\|; \\ p^{\sigma(j)} & , \text{ otherwise.} \end{cases}$$

By combining (23), Corollary 6.1 and Lemma 3.1 it follows that

$$\begin{aligned}\Lambda &\geq \frac{1}{2} \sum_{j=1}^k \sum_{u \in A_j \cap P_{\gamma(j)}} d_u \|p^{\gamma(j)} - c_j\|^2 - \sum_{j=1}^k \sum_{u \in A_j \cap P_{\gamma(j)}} d_u \|F(u) - p^{\gamma(j)}\|^2 \\ &\geq \frac{1}{16} \sum_{j=1}^k \frac{\mu(A_j \cap P_{\gamma(j)})}{\min\{\mu(P_{\sigma(j)}), \mu(P_\ell)\}} - \left(1 + \frac{3k}{\Psi}\right) \cdot \frac{k^2}{\Psi} \geq \frac{1}{16} - \frac{2k^2}{\Psi}.\end{aligned}$$

■

We are now ready to prove Lemma 2.4.

Proof of Lemma 2.4. We apply Lemma 6.2 with $\varepsilon' = \varepsilon/k$. Then by Lemma 6.3 we have

$$\text{Cost}(\{A_i, c_i\}_{i=1}^k) \geq \frac{\varepsilon}{16k} - \frac{2k^2}{\Psi},$$

and the desired result follows by setting $\varepsilon \geq 64\alpha \cdot k^3/\Psi$.

■

7 The Normalized Spectral Embedding is ε -separated

In this section, we prove that the normalized spectral embedding \mathcal{X}_V is ε -separated.

Proof of Theorem 2.6 We establish first a lower bound on $\Delta_{k-1}(\mathcal{X}_V)$.

Lemma 7.1. *Let G be a graph that satisfies $\Psi = 20^4 \cdot k^3/\delta$ for some $\delta \in (0, 1/2]$. Then for $\delta' = 2\delta/20^4$ it holds*

$$\Delta_{k-1}(\mathcal{X}_V) \geq \frac{1}{12} - \frac{\delta'}{k}. \quad (25)$$

Before we present the proof of Lemma 7.1 we show that it implies (11). By Lemma 3.1 we have

$$\Delta_k(\mathcal{X}_V) \leq \frac{2k^2}{\Psi} = \frac{\delta'}{k}.$$

Moreover, by applying Lemma 7.1 with $k/\delta \geq 10^9$ and $\varepsilon = 6 \cdot 10^{-7}$ we obtain

$$\Delta_{k-1}(\mathcal{X}_V) \geq \frac{1}{12} - \frac{\delta'}{k} = \frac{1}{12} - \frac{2}{20^4} \cdot \frac{\delta}{k} \geq \frac{10^{10}}{9 \cdot 2^5} \cdot \frac{\delta}{k} = \frac{1}{\varepsilon^2} \cdot \frac{\delta'}{k} \geq \frac{1}{\varepsilon^2} \cdot \Delta_k(\mathcal{X}_V).$$

Proof of Lemma 7.1 We argue in a similar manner as in Lemma 6.3 (c.f. Case 3). We start by giving some notation, then we establish Lemma 7.2 and apply it in the proof of Lemma 7.1.

We redefine the function σ (c.f. (20)) such that for any two partitions (P_1, \dots, P_k) and (Z_1, \dots, Z_{k-1}) of V , we define a mapping $\sigma : [1 : k-1] \mapsto [1 : k]$ by

$$\sigma(i) = \arg \max_{j \in [1:k]} \frac{\mu(Z_i \cap P_j)}{\mu(P_j)}, \quad \text{for every } i \in [1 : k-1].$$

We lower bound now the clusters overlapping in terms of the volume between any k -way and $(k-1)$ -way partitions of V .

Lemma 7.2. *Suppose (P_1, \dots, P_k) and (Z_1, \dots, Z_{k-1}) are partitions of V . Then for any index $\ell \in [1 : k] \setminus \{\sigma(1), \dots, \sigma(k-1)\}$ (there is at least one such ℓ) and for every $i \in [1 : k-1]$ it holds*

$$\{\mu(Z_i \cap P_{\sigma(i)}), \mu(Z_i \cap P_\ell)\} \geq \tau_i \cdot \min\{\mu(P_\ell), \mu(P_{\sigma(i)})\},$$

where $\sum_{i=1}^{k-1} \tau_i = 1$ and $\tau_i \geq 0$.

Proof. By pigeonhole principle there is an index $\ell \in [1 : k]$ such that $\ell \notin \{\sigma(1), \dots, \sigma(k-1)\}$. Thus, for every $i \in [1 : k-1]$ we have $\sigma(i) \neq \ell$ and

$$\frac{\mu(Z_i \cap P_{\sigma(i)})}{\mu(P_{\sigma(i)})} \geq \frac{\mu(Z_i \cap P_\ell)}{\mu(P_\ell)} \triangleq \tau_i,$$

where $\sum_{i=1}^{k-1} \tau_i = 1$ and $\tau_i \geq 0$ for all i . Hence, the statement follows. ■

We present now the proof of Lemma 7.1.

Proof of Lemma 7.1. Let (Z_1, \dots, Z_{k-1}) be a $(k-1)$ -way partition of V with centers c'_1, \dots, c'_{k-1} that achieves $\Delta_{k-1}(\mathcal{X}_V)$, and (P_1, \dots, P_k) be a k -way partition of V achieving $\hat{\rho}_{\text{avt}}(k)$. Our goal now is to lower bound the optimal $(k-1)$ -means cost

$$\Delta_{k-1}(\mathcal{X}_V) = \sum_{i=1}^{k-1} \sum_{j=1}^k \sum_{u \in Z_i \cap P_j} d_u \|F(u) - c'_i\|^2. \quad (26)$$

By Lemma 7.2 there is an index $\ell \in [1 : k] \setminus \{\sigma(1), \dots, \sigma(k-1)\}$. For $i \in [1 : k-1]$ let

$$p^{\gamma(i)} = \begin{cases} p^\ell & , \text{ if } \|p^\ell - c'_i\| \geq \|p^{\sigma(i)} - c'_i\|; \\ p^{\sigma(i)} & , \text{ otherwise.} \end{cases}$$

Then by combining Corollary 6.1 and Lemma 7.2, we have

$$\|p^{\gamma(i)} - c'_i\|^2 \geq [8 \cdot \min\{\mu(P_\ell), \mu(P_{\sigma(i)})\}]^{-1} \text{ and } \mu(Z_i \cap P_{\gamma(i)}) \geq \tau_i \cdot \min\{\mu(P_\ell), \mu(P_{\sigma(i)})\}, \quad (27)$$

where $\sum_{i=1}^{k-1} \tau_i = 1$. We now lower bound the expression in (26). Since

$$\|F(u) - c'_i\|^2 \geq \frac{1}{2} \|p^{\gamma(i)} - c'_i\|^2 - \|F(u) - p^{\gamma(i)}\|^2,$$

it follows for $\delta' = 2\delta/20^4$ that

$$\begin{aligned}
\triangle_{k-1}(\mathcal{X}_V) &= \sum_{i=1}^{k-1} \sum_{j=1}^k \sum_{u \in Z_i \cap P_j} d_u \|F(u) - c'_i\|^2 \geq \sum_{i=1}^{k-1} \sum_{u \in Z_i \cap P_{\gamma(i)}} d_u \|F(u) - c'_i\|^2 \\
&\geq \frac{1}{2} \sum_{i=1}^{k-1} \sum_{u \in Z_i \cap P_{\gamma(i)}} d_u \|p^{\gamma(i)} - c'_i\|^2 - \sum_{i=1}^{k-1} \sum_{u \in Z_i \cap P_{\gamma(i)}} d_u \|F(u) - p^{\gamma(i)}\|^2 \\
&\geq \frac{1}{2} \sum_{i=1}^{k-1} \frac{\mu(Z_i \cap P_{\gamma(i)})}{8 \cdot \min\{\mu(P_{\gamma(i)}), \mu(P_{\sigma(i)})\}} - \sum_{i=1}^k \sum_{u \in P_i} d_u \|F(u) - p^i\|^2 \\
&\geq \frac{1}{16} - \frac{\delta'}{k},
\end{aligned}$$

where the last inequality holds due to (27) and Lemma 3.1. \blacksquare

8 An Efficient Spectral Clustering Algorithm

In this section, we apply the proof techniques developed by Boutsidis et al. [1, 2] to our setting. More precisely, we prove that any α -approximate k -means algorithm that runs on an approximate normalized spectral embedding \widetilde{Y}' computed by the power method, yields an approximate clustering \widetilde{X}'_α of the normalized spectral embedding Y' .

Furthermore, we prove under our gap assumption that \widetilde{Y}' is ε -separated. This allows us to apply the variant of Lloyd's k -means algorithm analyzed by Ostrovsky et al. [7] to efficiently compute \widetilde{X}'_α . Then we use part (a) of Theorem 1.2 to establish the desired statement.

This Section is organized as follows. In Subsection 8.1, we prove Lemma 2.7. In Subsection 8.2, we present the proof of Theorem 2.8. In Subsection 8.3, we establish Theorem 2.9. Based on the results from the preceding three subsections, we prove part (b) of Theorem 1.2 in Subsection 8.4.

8.1 Proof of Lemma 2.7

We argue in a similar manner as in [1, Lemma 7]. By the eigenvalue decomposition theorem $\mathcal{B} = U\Sigma U^T$, where the columns of $U = [U_k \ U_{\rho-k}] \in \mathbb{R}^{n \times \rho}$ are the orthonormal eigenbasis of \mathcal{L}_G , $\Sigma \in \mathbb{R}^{\rho \times \rho}$ is a positive diagonal matrix such that $\Sigma_{ii} = 2 - \lambda_i \geq 0$ for all i , and $\rho = \text{rank}(\mathcal{B})$.

Since $\mathcal{B}^p = U\Sigma^p U^T$, it follows that $\ker(\mathcal{B}^p S) = \ker(U^T S)$. By Lemma A.3 with probability at least $1 - e^{-2n}$ we have $\text{rank}(U^T S) = k$ and thus matrix $\mathcal{B}^p S$ has k singular values. Furthermore, the SVD decomposition $\widetilde{U}\widetilde{\Sigma}\widetilde{V}^T$ of $\mathcal{B}^p S$ satisfies: $\widetilde{U} \in \mathbb{R}^{n \times k}$ is a matrix with orthonormal columns, $\widetilde{\Sigma} \in \mathbb{R}^{k \times k}$ is a positive diagonal matrix and $\widetilde{V}^T \in \mathbb{R}^{k \times k}$ is an orthonormal matrix. We denote by

$$R \triangleq \widetilde{\Sigma}\widetilde{V}^T \in \mathbb{R}^{k \times k},$$

and we use the following four facts

$$\widetilde{U}R = U_k \Sigma_k^p U_k^T S + U_{\rho-k} \Sigma_{\rho-k}^p U_{\rho-k}^T S \quad (28)$$

$$\sigma_i(\widetilde{U}R) \geq \sigma_k(U_k \Sigma_k^p U_k^T S) \geq (2 - \lambda_k)^p \cdot \sigma_k(U_k^T S) \quad (29)$$

$$\sigma_i(\widetilde{U}R) = \sigma_i(R) \quad (30)$$

$$\|X\widetilde{U}\|_2 \geq \|X\widetilde{U}\|_2 \cdot \sigma_k(\widetilde{U}), \quad \text{for any } X \in \mathbb{R}^{\ell \times k} \quad (31)$$

(28) follows from the eigenvalue decomposition of \mathcal{B} and the fact that $\mathcal{B}^p = U\Sigma^p U^T$; (29) follows by (28) due to U_k and $U_{\rho-k}$ span orthogonal spaces, and since the minimum singular value of a product is at least the product of the minimum singular values; (30) holds due to $\tilde{U}^T \tilde{U} = I_k$; Recall that with probability at least $1 - e^{-2n}$ we have $\sigma_k(R) > 0$ and hence (31) follows by

$$\|X\|_2 = \max_{x \neq 0} \frac{\|XRx\|_2}{\|Rx\|_2} \leq \max_{x \neq 0} \frac{\|XRx\|_2}{\sigma_k(R) \|x\|_2} = \frac{\|XR\|_2}{\sigma_k(R)}.$$

[4, Theorem 2.6.1] shows that for every two $m \times k$ orthonormal matrices W, Z with $m \geq k$ it holds

$$\|WW^T - ZZ^T\|_2 = \|Z^T W^\perp\|_2 = \|W^T Z^\perp\|_2,$$

where $[Z, Z^\perp] \in \mathbb{R}^{m \times m}$ is full orthonormal basis. Therefore, we have

$$\|U_k U_k^T - \tilde{U} \tilde{U}^T\|_2 = \|\tilde{U}^T U_k^\perp\|_2 = \left\| \left(U_k^\perp \right)^T \tilde{U} \right\|_2 = \|U_{\rho-k}^T \tilde{U}\|_2, \quad (32)$$

where the last equality is due to $U_{n-\rho}^T \tilde{U} = 0$, since \tilde{U} is in the range of $U = U_\rho$.

To upper bound $\|U_{\rho-k}^T \tilde{U}\|_2$ we establish the next two inequalities:

$$\|U_{\rho-k}^T \tilde{U} R\|_2 \geq \|U_{\rho-k}^T \tilde{U}\|_2 \cdot \sigma(R) \geq \|U_{\rho-k}^T \tilde{U}\|_2 \cdot (2 - \lambda_k)^p \cdot \sigma_k(U_k^T S), \quad (33)$$

$$\|U_{\rho-k}^T \tilde{U} R\|_2 = \|\Sigma_{\rho-k}^p U_{\rho-k}^T S\|_2 \leq (2 - \lambda_{k+1})^p \cdot \sigma_1(U_{\rho-k}^T S), \quad (34)$$

where (33) follows by (31), (30) and (29); and (34) is due to (28) and $\Sigma_{11} \geq \dots \geq \Sigma_{nn} \geq 0$.

By combining Lemma A.1 and Lemma A.2 with probability at least $1 - e^{-2n} - 3\delta$ both inequalities hold

$$\frac{\delta}{\sqrt{k}} \leq \sigma_k(U_k^T S) \quad \text{and} \quad \sigma_1(U_{\rho-k}^T S) \leq 4\sqrt{n}. \quad (35)$$

Using (32), (33), (34) and (35) we obtain

$$\|U_k U_k^T - \tilde{U} \tilde{U}^T\|_2 = \|U_{\rho-k}^T \tilde{U}\|_2 \leq 4\delta^{-1} \cdot \sqrt{nk} \cdot \gamma_k^p. \quad (36)$$

Since $\|M\|_F \leq \sqrt{\text{rank}(M)} \cdot \|M\|_2$ for every matrix M and $\text{rank}(U_k U_k^T - \tilde{U} \tilde{U}^T) \leq 2k$, it follows

$$\|U_k U_k^T - \tilde{U} \tilde{U}^T\|_F \leq 2k \cdot \|U_k U_k^T - \tilde{U} \tilde{U}^T\|_2 \leq 8\delta^{-1} \cdot n^{1/2} k^{3/2} \cdot \gamma_k^p \leq \epsilon,$$

where the last two inequalities are due to (36) and the choice of γ_k .

8.2 Proof of Theorem 2.8

We establish several technical Lemmas that combined with Lemma 2.7 allow us to apply the proof techniques in [2, Theorem 6].

Lemma 8.1. $X' X'^T$ is a projection matrix.

Proof. By construction there are $d(v)$ many copies of row $U_k(v, :)/\sqrt{d(v)}$ in Y' , for every vertex $v \in V$. We may assume w.l.o.g. that a k -means algorithm outputs an indicator matrix X' such that all copies of row $U_k(v, :)/\sqrt{d(v)}$ belong to the same cluster, for every $v \in V$. Moreover, by definition $X'_{ij} = 1/\sqrt{\mu(C_j)}$ if row $Y'_{i,:}$ belongs to the j -th cluster C_j and $X'_{ij} = 0$ otherwise, where matrix $X' \in \mathbb{R}^{m \times k}$. Therefore, it follows that $X'^T X' = I_{k \times k}$ and thus $(X' X'^T)^2 = X' X'^T$. ■

Lemma 8.2. *It holds that $Y'^T Y' = I_{k \times k} = \widetilde{Y}'^T \widetilde{Y}'$.*

Proof. We prove now $Y'^T Y' = I_{k \times k}$, but the equality $\widetilde{Y}'^T \widetilde{Y}' = I_{k \times k}$ follows similarly. Since

$$\begin{aligned} (Y'^T Y')_{ij} &= \left(\frac{U_k(1,i)}{\sqrt{d(1)}} \mathbf{1}_{d(1)}^T \quad \cdots \quad \frac{U_k(n,i)}{\sqrt{d(n)}} \mathbf{1}_{d(n)}^T \right) \begin{pmatrix} \frac{U_k(1,j)}{\sqrt{d(1)}} \mathbf{1}_{d(1)} \\ \vdots \\ \frac{U_k(n,j)}{\sqrt{d(n)}} \mathbf{1}_{d(n)} \end{pmatrix} \\ &= \sum_{\ell=1}^n d(\ell) \frac{U_k(\ell,i)}{\sqrt{d(\ell)}} \frac{U_k(\ell,j)}{\sqrt{d(\ell)}} = \langle U_k(:,i), U_k(:,j) \rangle = \delta_{ij}, \end{aligned}$$

the statement follows. ■

Lemma 8.3. *It holds that $\|Y' Y'^T - \widetilde{Y}' \widetilde{Y}'^T\|_F = \|Y Y^T - \widetilde{Y} \widetilde{Y}^T\|_F$.*

Proof. By definition

$$Y' Y'^T = \sum_{\ell=1}^k Y'_{:, \ell} Y'^T_{:, \ell} \quad \text{where} \quad Y'_{:, \ell} = \begin{pmatrix} \frac{U_k(1,\ell)}{\sqrt{d(1)}} \mathbf{1}_{d(1)} \\ \vdots \\ \frac{U_k(n,\ell)}{\sqrt{d(n)}} \mathbf{1}_{d(n)} \end{pmatrix}_{m \times 1}$$

and

$$(Y'_{:, \ell} Y'^T_{:, \ell})_{d(i)d(j)} = \frac{U_k(i,\ell) U_k(j,\ell)}{\sqrt{d(i)d(j)}} \cdot \mathbf{1}_{d(1)} \mathbf{1}_{d(j)}^T.$$

The statement follows by establishing the following chain of equalities

$$\begin{aligned} \|Y' Y'^T - \widetilde{Y}' \widetilde{Y}'^T\|_F^2 &= \sum_{i=1}^n \sum_{j=1}^n \left\| (Y' Y'^T - \widetilde{Y}' \widetilde{Y}'^T)_{d(i)d(j)} \right\|_F^2 \\ &= \sum_{i=1}^n \sum_{j=1}^n \left\| \sum_{\ell=1}^k (Y'_{:, \ell} Y'^T_{:, \ell} - \widetilde{Y}'_{:, \ell} \widetilde{Y}'^T_{:, \ell})_{d(i)d(j)} \right\|_F^2 \\ &= \sum_{i=1}^n \sum_{j=1}^n \left\| \left\{ \sum_{\ell=1}^k \left(\frac{U_k(i,\ell) U_k(j,\ell)}{\sqrt{d(i)d(j)}} - \frac{\widetilde{U}(i,\ell) \widetilde{U}(j,\ell)}{\sqrt{d(i)d(j)}} \right) \right\} \cdot \mathbf{1}_{d(i)} \mathbf{1}_{d(j)}^T \right\|_F^2 \\ &= \sum_{i=1}^n \sum_{j=1}^n d(i)d(j) \left[\sum_{\ell=1}^k \left(\frac{U_k(i,\ell) U_k(j,\ell)}{\sqrt{d(i)d(j)}} - \frac{\widetilde{U}(i,\ell) \widetilde{U}(j,\ell)}{\sqrt{d(i)d(j)}} \right) \right]^2 \\ &= \sum_{i=1}^n \sum_{j=1}^n \left[\sum_{\ell=1}^k (U_k(i,\ell) U_k(j,\ell) - \widetilde{U}(i,\ell) \widetilde{U}(j,\ell)) \right]^2 \\ &= \sum_{i=1}^n \sum_{j=1}^n (U_k U_k^T - \widetilde{U} \widetilde{U}^T)_{ij}^2 \\ &= \|U_k U_k^T - \widetilde{U} \widetilde{U}^T\|_F^2 = \|Y Y^T - \widetilde{Y} \widetilde{Y}^T\|_F^2. \end{aligned}$$
■

Lemma 8.4. *For any matrix U with orthonormal columns and every matrix A it holds*

$$\|UU^T - AA^TUU^T\|_F = \|U - AA^TU\|_F. \quad (37)$$

Proof. The statement follows by the Frobenius norm property $\|M\|_F^2 = \text{Tr}[M^T M]$, the cyclic property of trace $\text{Tr}[UM^T MU^T] = \text{Tr}[M^T M \cdot U^T U]$ and the orthogonality of matrix U . ■

We are now ready to prove Theorem 2.8.

Proof of Theorem 2.8. Using Lemma 2.7 and Lemma 8.3 with probability at least $1 - 2e^{-2n} - 3\delta_p$ we have

$$\|Y'Y'^T - \widetilde{Y}'\widetilde{Y}'^T\|_F = \|YY^T - \widetilde{Y}\widetilde{Y}^T\|_F \leq \varepsilon.$$

Let $Y'Y'^T = \widetilde{Y}'\widetilde{Y}'^T + E$ such that $\|E\|_F \leq \varepsilon$. By combining Lemma 8.2 and Lemma 8.4, (37) holds for the matrices Y' and \widetilde{Y}' . Thus, by Lemma 8.1 and the proof techniques in [2, Theorem 6] it follows

$$\left\|Y' - \widetilde{X}'_\alpha \left(\widetilde{X}'_\alpha\right)^T Y'\right\|_F \leq \sqrt{\alpha} \cdot \left(\left\|Y' - X'_{\text{opt}} \left(X'_{\text{opt}}\right)^T Y'\right\|_F + 2\varepsilon\right). \quad (38)$$

The desired statement follows by simple algebraic manipulations of (38). ■

8.3 Proof of Theorem 2.9

In this subsection, we show under our gap assumption that the approximate normalized spectral embedding \widetilde{Y}' is ε -separated, i.e. $\Delta_k(\mathcal{X}_V) < 5\varepsilon^2 \cdot \Delta_{k-1}(\mathcal{X}_V)$. Our analysis builds upon Theorem 2.6, Theorem 2.8 and the proof techniques in [2, Theorem 6].

Before we present the proof of Theorem 2.9 we establish two technical Lemmas.

Lemma 8.5. *If $\Psi \geq 20^4 \cdot k^3/\delta$ for $\delta \in (0, 1/2]$ it holds*

$$\ln\left(\frac{2 - \lambda_k}{2 - \lambda_{k+1}}\right) \geq \frac{1}{2} \left(1 - \frac{4\delta}{20^4 k^2}\right) \lambda_{k+1}.$$

Proof. Lee et al. [5] proved that higher order Cheeger's inequality satisfies

$$\lambda_k/2 \leq \rho(k) \leq O(k^2) \cdot \sqrt{\lambda_k}. \quad (39)$$

Using the LHS of (39) we have

$$k^3 \widehat{\rho}_{\text{avr}}(k) = k^2 \sum_{i=1}^k \phi(P_i) \geq k^2 \max_{i \in [1:k]} \phi(P_i) \geq k^2 \cdot \rho(k) \geq \frac{k^2 \lambda_k}{2}$$

and thus we can upper bound the k -th smallest eigenvalue of \mathcal{L}_G by

$$\lambda_k \leq 2k \cdot \widehat{\rho}_{\text{avr}}(k).$$

Moreover, by the gap assumption we have

$$\lambda_{k+1} \geq \frac{20^4 k^2}{2\delta} \cdot 2k \cdot \widehat{\rho}_{\text{avr}}(k) \geq \frac{20^4 k^2}{2\delta} \cdot \lambda_k.$$

The statement follows by

$$\frac{2 - \lambda_k}{2 - \lambda_{k+1}} \geq \frac{1 - \frac{\delta}{20^4 k^2} \cdot \lambda_{k+1}}{1 - \frac{1}{2} \lambda_{k+1}} \geq \exp\left\{\frac{1}{2} \left(1 - \frac{4\delta}{20^4 k^2}\right) \lambda_{k+1}\right\}.$$

■

For our next results we need to introduce some notation. We use interchangeably X'_{opt} with $X_{\text{opt}}^{(k)}$ to denote the optimal indicator matrix for the k -means problem on \mathcal{X}_V that is induced by the rows of matrix Y' . Similarly, we denote by $X_{\text{opt}}^{(k-1)}$ the optimal indicator matrix for the $(k-1)$ -means problem on \mathcal{X}_V .

Based on Lemma 3.1 and the definition of Y' and $X_{\text{opt}}^{(k)}$ we obtain the following statement.

Corollary 8.6. *Let G be a graph that satisfies $\Psi = 20^4 \cdot k^3/\delta$, $\delta \in (0, 1/2]$ and $k/\delta \geq 10^9$. Then it holds*

$$\left\| Y' - X_{\text{opt}}^{(k)} \left(X_{\text{opt}}^{(k)} \right)^T Y' \right\|_F^2 \leq \frac{1}{8 \cdot 10^{13}}.$$

We are now ready to prove Theorem 2.9.

Proof of Theorem 2.9. By Theorem 2.6 we have

$$\left\| Y' - X_{\text{opt}}^{(k)} \left(X_{\text{opt}}^{(k)} \right)^T Y' \right\|_F \leq \varepsilon \left\| Y' - X_{\text{opt}}^{(k-1)} \left(X_{\text{opt}}^{(k-1)} \right)^T Y' \right\|_F. \quad (40)$$

We set the approximation parameter in Theorem 2.8 to

$$\varepsilon' \triangleq \frac{1}{4} \sqrt{\Delta_k(\mathcal{X}_V)} = \frac{1}{4} \left\| Y' - X_{\text{opt}}^{(k)} \left(X_{\text{opt}}^{(k)} \right)^T Y' \right\|_F \geq n^{-O(1)}, \quad (41)$$

and we note that by Theorem 2.6 it holds

$$\varepsilon' \leq \frac{\varepsilon}{4} \sqrt{\Delta_{k-1}(\mathcal{X}_V)}. \quad (42)$$

We construct now matrix \tilde{Y} via the power method with $p \geq \Omega(\frac{\ln n}{\lambda_{k+1}})$. By combining Lemma 2.7 and Lemma 8.3 we obtain with high probability that

$$\left\| Y' Y'^T - \tilde{Y} \tilde{Y}^T \right\|_F = \left\| Y Y^T - \tilde{Y} \tilde{Y}^T \right\|_F \leq \varepsilon'.$$

Let $Y' Y'^T = \tilde{Y} \tilde{Y}^T + E$ such that $\|E\|_F \leq \varepsilon'$. By Lemma 8.2 we have $Y'^T Y' = I_{k \times k} = \tilde{Y}^T \tilde{Y}$ and thus (37) in Lemma 8.4 holds for the orthonormal matrices Y' and \tilde{Y} . Therefore, by Lemma 8.1 we have

$$\begin{aligned} \sqrt{\Delta_k(\mathcal{X}_V)} &= \left\| \tilde{Y}' - \widetilde{X_{\text{opt}}^{(k)}} \left(\widetilde{X_{\text{opt}}^{(k)}} \right)^T \tilde{Y}' \right\|_F = \left\| \tilde{Y}' \tilde{Y}'^T - \widetilde{X_{\text{opt}}^{(k)}} \left(\widetilde{X_{\text{opt}}^{(k)}} \right)^T \tilde{Y}' \tilde{Y}'^T \right\|_F \\ &= \left\| Y' Y'^T - \widetilde{X_{\text{opt}}^{(k)}} \left(\widetilde{X_{\text{opt}}^{(k)}} \right)^T Y' Y'^T - \left(I - \widetilde{X_{\text{opt}}^{(k)}} \left(\widetilde{X_{\text{opt}}^{(k)}} \right)^T \right) E \right\|_F \\ &\leq \|E\|_F + \left\| Y' - \widetilde{X_{\text{opt}}^{(k)}} \left(\widetilde{X_{\text{opt}}^{(k)}} \right)^T Y' \right\|_F \end{aligned}$$

By Lemma 8.5 we can apply Theorem 2.8 which yields

$$\left\| Y' - \widetilde{X_{\text{opt}}^{(k)}} \left(\widetilde{X_{\text{opt}}^{(k)}} \right)^T Y' \right\|_F^2 \leq (1 + 4\varepsilon') \cdot \left\| Y' - X_{\text{opt}}^{(k)} \left(X_{\text{opt}}^{(k)} \right)^T Y' \right\|_F^2 + 4\varepsilon'^2.$$

By Corollary 8.6 we derive an upper bound on the optimal k -means cost of \mathcal{X}_V

$$\left\| Y' - X_{\text{opt}}'^{(k)} \left(X_{\text{opt}}'^{(k)} \right)^T Y' \right\|_F^2 \leq \frac{1}{8 \cdot 10^{13}} \quad (43)$$

that combined with the definition of ε' gives

$$\begin{aligned} \sqrt{\Delta_k(\widetilde{\mathcal{X}}_V)} &\leq \varepsilon' + \sqrt{(1 + 4\varepsilon') \left\| Y' - X_{\text{opt}}'^{(k)} \left(X_{\text{opt}}'^{(k)} \right)^T Y' \right\|_F^2 + 4\varepsilon'^2} \\ &\leq 2 \left\| Y' - X_{\text{opt}}'^{(k)} \left(X_{\text{opt}}'^{(k)} \right)^T Y' \right\|_F = 2\sqrt{\Delta_k(\mathcal{X}_V)} \\ &\leq 2\varepsilon \cdot \sqrt{\Delta_{k-1}(\mathcal{X}_V)}. \end{aligned} \quad (44)$$

Moreover, it holds that

$$\begin{aligned} \sqrt{\Delta_{k-1}(\mathcal{X}_V)} &= \left\| Y' - X_{\text{opt}}'^{(k-1)} \left(X_{\text{opt}}'^{(k-1)} \right)^T Y' \right\|_F \leq \left\| Y' - \widetilde{X_{\text{opt}}'^{(k-1)}} \left(\widetilde{X_{\text{opt}}'^{(k-1)}} \right)^T Y' \right\|_F \\ &= \left\| Y' Y'^T - \widetilde{X_{\text{opt}}'^{(k-1)}} \left(\widetilde{X_{\text{opt}}'^{(k-1)}} \right)^T Y' Y'^T \right\|_F \\ &= \left\| \widetilde{Y' Y'^T} - \widetilde{X_{\text{opt}}'^{(k-1)}} \left(\widetilde{X_{\text{opt}}'^{(k-1)}} \right)^T \widetilde{Y' Y'^T} + \left(I - \widetilde{X_{\text{opt}}'^{(k-1)}} \left(\widetilde{X_{\text{opt}}'^{(k-1)}} \right)^T \right) E \right\|_F \\ &\leq \left\| \widetilde{Y'} - \widetilde{X_{\text{opt}}'^{(k-1)}} \left(\widetilde{X_{\text{opt}}'^{(k-1)}} \right)^T \widetilde{Y'} \right\|_F + \|E\|_F \\ &\leq \sqrt{\Delta_{k-1}(\widetilde{\mathcal{X}}_V)} + \frac{\varepsilon}{4} \sqrt{\Delta_{k-1}(\mathcal{X}_V)} \end{aligned}$$

and thus

$$\sqrt{\Delta_{k-1}(\mathcal{X}_V)} \leq \left(1 + \frac{\varepsilon}{2}\right) \sqrt{\Delta_{k-1}(\widetilde{\mathcal{X}}_V)}. \quad (45)$$

The statement follows by combining (44) and (45).

$$\sqrt{\Delta_k(\widetilde{\mathcal{X}}_V)} \leq 2\varepsilon \cdot \sqrt{\Delta_{k-1}(\mathcal{X}_V)} \leq (2 + \varepsilon) \cdot \varepsilon \cdot \sqrt{\Delta_{k-1}(\widetilde{\mathcal{X}}_V)}.$$

■

8.4 Proof of Part (b) of Theorem 1.2

Let $p = \Theta(\frac{\ln n}{\lambda_{k+1}})$. We compute the matrix $\mathcal{B}^p S$ in time $O(mkp)$ and its singular value decomposition $\widetilde{U} \widetilde{\Sigma} \widetilde{V}^T$ in time $O(nk^2)$. Based on it, we construct in time $O(mk)$ matrix $\widetilde{Y'}$ (c.f. (14)).

By Theorem 2.9, $\widetilde{\mathcal{X}}_V$ is ε -separated for $\varepsilon = 6 \cdot 10^{-7}$, i.e. $\Delta_k(\widetilde{\mathcal{X}}_V) < 5\varepsilon^2 \cdot \Delta_{k-1}(\widetilde{\mathcal{X}}_V)$. Hence, by Theorem 2.5 there is an algorithm that outputs a clustering with indicator matrix \widetilde{X}'_α that has a cost at most

$$\left\| \widetilde{Y'} - \widetilde{X}'_\alpha \left(\widetilde{X}'_\alpha \right)^T \widetilde{Y'} \right\|_F^2 \leq \left(1 + \frac{1}{10^{10}}\right) \cdot \left\| \widetilde{Y'} - \widetilde{X}'_{\text{opt}} \left(\widetilde{X}'_{\text{opt}} \right)^T \widetilde{Y'} \right\|_F^2$$

with constant probability (close to 1) in time $O(mk^2 + k^4)$, where $\alpha = 1 + 10^{-10}$.

We apply Theorem 2.8 with $\varepsilon' = \frac{\sqrt{\delta_A}}{4} \|Y' - X'_{\text{opt}} (X'_{\text{opt}})^T Y'\|_F$, where $\delta_A \in (0, 1)$ is to be determined soon, and since by Corollary 8.6 we have

$$\|Y' - X'_{\text{opt}} (X'_{\text{opt}})^T Y'\|_F < \frac{1}{10^6},$$

with constant probability it follows that

$$\begin{aligned} & \left\| Y' - \widetilde{X}'_{\alpha} \left(\widetilde{X}'_{\alpha} \right)^T Y' \right\|_F^2 \leq (1 + 4\varepsilon') \alpha \left\| Y' - X'_{\text{opt}} (X'_{\text{opt}})^T Y' \right\|_F^2 + 4\varepsilon'^2 \\ &= \left[\left(1 + \sqrt{\delta_A} \left\| Y' - X'_{\text{opt}} (X'_{\text{opt}})^T Y' \right\|_F \right) \alpha + \frac{\delta_A}{4} \right] \cdot \left\| Y' - X'_{\text{opt}} (X'_{\text{opt}})^T Y' \right\|_F^2 \\ &\leq \left[\left(1 + \frac{\sqrt{\delta_A}}{10^6} \right) \cdot \left(1 + \frac{1}{10^{10}} \right) + \frac{\delta_A}{4} \right] \cdot \left\| Y' - X'_{\text{opt}} (X'_{\text{opt}})^T Y' \right\|_F^2. \end{aligned}$$

The indicator matrix \widetilde{X}'_{α} yields a multiplicative approximation of \mathcal{X}_V that satisfies for $\delta_A = 1/10^6$

$$\left\| Y' - \widetilde{X}'_{\alpha} \left(\widetilde{X}'_{\alpha} \right)^T Y' \right\|_F^2 \leq \left(1 + \frac{1}{10^6} \right) \left\| Y' - X'_{\text{opt}} (X'_{\text{opt}})^T Y' \right\|_F^2. \quad (46)$$

The statement follows by part (a) of Theorem 1.2 applied to the partition (A_1, \dots, A_k) of V that is induced by the indicator matrix \widetilde{X}'_{α} .

9 Parameterized Upper Bound on $\widehat{\rho}_{\text{avr}}(k)$

A k -disjoint tuple Z is a k -tuple (Z_1, \dots, Z_k) of disjoint subsets of V . A k -way partition (P_1, \dots, P_k) of V is compatible with a k -disjoint tuple Z if $Z_i \subseteq P_i$ for all i . We then define $S_i = P_i \setminus Z_i$ and use \mathcal{P}_Z to denote all partitions compatible with Z . We use \mathcal{Z}_k to denote all k -tuples Z with $\rho(k) = \Phi(Z) = \Phi(Z_1, \dots, Z_k)$. The elements of \mathcal{Z}_k are called optimal (k -disjoint) tuples. We denote all partitions compatible with some optimal k -tuple by

$$\mathcal{P}_k = \cup_{Z \in \mathcal{Z}_k} \mathcal{P}_Z. \quad (47)$$

Oveis Gharan and Trevisan [8, Lemma 2.5] proved that for every k -disjoint tuple $Z \in \mathcal{Z}_k$ there is a k -way partition $(P_1, \dots, P_k) \in \mathcal{P}_Z$ with

$$\Phi(P_1, \dots, P_k) \leq k\rho(k). \quad (48)$$

Remark 9.1. In this section, we assume that every partition $(P_1, \dots, P_k) \in \mathcal{P}_k$ satisfies

$$\Phi(P_1, \dots, P_k) > \rho(k), \quad (49)$$

since otherwise $\widehat{\rho}(k) = \rho(k)$.

We refine the analysis in [8] and prove a parameterized upper bound on $\widehat{\rho}_{\text{avr}}(k)$ that depends on a natural combinatorial parameter and the average conductance of a k -disjoint tuple $Z \in \mathcal{Z}_k$. Before we state our results, we need some notation.

We define the order k inter-connection constant of a graph G by

$$\rho_{\mathcal{P}}(k) \triangleq \min_{P_1, \dots, P_k \in \mathcal{P}_k} \Phi_{IC}(P_1, \dots, P_k) \quad (50)$$

where

$$\Phi_{IC}(P_1, \dots, P_k) \triangleq \max_{S_i \neq \emptyset} \frac{|E(S_i, V \setminus P_i)| - |E(S_i, Z_i)|}{|E(P_i, V \setminus P_i)|}. \quad (51)$$

We will prove in Lemma 9.5 that $\rho_{\mathcal{P}}(k) \in (0, 1 - 1/(k-1)]$. Furthermore, let $\mathcal{O}_{\mathcal{P}}$ be the set of all k -way partitions $(P_1, \dots, P_k) \in \mathcal{P}_k$ with $\Phi_{IC}(P_1, \dots, P_k) = \rho_{\mathcal{P}}(k)$, i.e., the set of all partitions that achieve the order k inter-connection constant. Let

$$\tilde{\rho}_{\text{avr}}(k) = \min_{(P_1, \dots, P_k) \in \mathcal{O}_{\mathcal{P}}} \frac{1}{k} \sum_{i=1}^k \phi(P_i) \quad (52)$$

be the minimal *average* conductance over all k -way partitions in $\mathcal{O}_{\mathcal{P}}$. By construction it holds that

$$\hat{\rho}_{\text{avr}}(k) \leq \tilde{\rho}_{\text{avr}}(k). \quad (53)$$

We present now our main result of this Section which upper bounds $\tilde{\rho}_{\text{avr}}(k)$.

Theorem 9.2. *For any graph G there exists a k -way partition $(P_1, \dots, P_k) \in \mathcal{O}_{\mathcal{P}}$ compatible with a k -disjoint tuple Z with $\Phi(Z_1, \dots, Z_k) = \rho(k)$ such that for $\kappa_{\mathcal{P}} \triangleq [1 - \rho_{\mathcal{P}}(k)]^{-1} \in (1, k-1]$ it holds*

$$\tilde{\rho}_{\text{avr}}(k) \leq \frac{\kappa_{\mathcal{P}}}{k} \sum_{i=1}^k \phi(Z_i)$$

and in addition, for every $i \in [1 : k]$

$$\phi(P_i) \leq \kappa_{\mathcal{P}} \cdot \phi(Z_i).$$

Our goal now is to prove Theorem 9.2. We establish first a few useful Lemmas that will be used to prove Lemma 9.5 and Theorem 9.2.

Oveis Gharan and Trevisan [8, Algorithm 2 and Fact 2.4] showed that

Fact 9.3 ([8]). *For any k -disjoint tuple Z , there is a k -way partition $(P_1, \dots, P_k) \in \mathcal{P}_Z$ such that*

1. *For every $i \in [1 : k]$, $Z_i \subseteq P_i$.*
2. *For every $i \in [1 : k]$, and every subset $\emptyset \neq S \subseteq P_i \setminus Z_i$ it holds*

$$|E(S, P_i \setminus S)| \geq \frac{1}{k} |E(S, V \setminus S)|.$$

Lemma 9.4. *For any k -disjoint tuple Z , there exists a k -way partition $(P_1, \dots, P_k) \in \mathcal{P}_Z$ that satisfies*

$$\max_{S_i \neq \emptyset} \frac{|E(S_i, V \setminus P_i)| - |E(S_i, Z_i)|}{|E(P_i, V \setminus P_i)|} \leq 1 - \frac{1}{k-1}.$$

Proof. By Fact 9.3 there is a k -way partition $(P_1, \dots, P_k) \in \mathcal{P}_Z$ such that for all i it holds

$$|E(S_i, Z_i)| = |E(S_i, P_i \setminus S_i)| \geq \frac{1}{k} |E(S_i, V \setminus S_i)| = \frac{1}{k} (|E(S_i, V \setminus P_i)| + |E(S_i, Z_i)|)$$

and hence

$$|E(S_i, Z_i)| \geq \frac{1}{k-1} |E(S_i, V \setminus P_i)|.$$

■

Lemma 9.5. *The order k inter-connection constant of a graph G is bounded by*

$$0 < \rho_{\mathcal{P}}(k) \leq 1 - \frac{1}{k-1}.$$

Proof. We prove first the upper bound. By Lemma 9.4 there is a k -way partition $(P_1, \dots, P_k) \in \mathcal{P}_k$ compatible with a k -disjoint tuple Z such that

$$\max_{S_i \neq \emptyset} \frac{|E(S_i, V \setminus P_i)| - |E(S_i, Z_i)|}{|E(P_i, V \setminus P_i)|} \leq 1 - \frac{1}{k-1}.$$

Therefore,

$$\begin{aligned} \rho_{\mathcal{P}}(k) &= \min_{P'_1, \dots, P'_k \in \mathcal{P}_k} \Phi_{IC}(P'_1, \dots, P'_k) \leq \Phi_{IC}(P_1, \dots, P_k) \\ &= \max_{S_i \neq \emptyset} \frac{|E(S_i, V \setminus P_i)| - |E(S_i, Z_i)|}{|E(P_i, V \setminus P_i)|} \leq 1 - \frac{1}{k-1}. \end{aligned}$$

We prove now the lower bound. Suppose for contradiction that $\rho_{\mathcal{P}}(k) \leq 0$. By definition we have

$$\begin{aligned} \phi(P_i) &= \frac{|E(P_i, V \setminus P_i)|}{\mu(P_i)} = \frac{|E(Z_i, V \setminus Z_i)| + |E(S_i, V \setminus P_i)| - |E(S_i, Z_i)|}{\mu(P_i)} \\ &\leq \phi(Z_i) + \frac{|E(S_i, V \setminus P_i)| - |E(S_i, Z_i)|}{\mu(P_i)} \end{aligned}$$

By (50), it holds for any $S_i \neq \emptyset$ that

$$|E(S_i, V \setminus P_i)| - |E(S_i, Z_i)| \leq \rho_{\mathcal{P}}(k) \cdot |E(P_i, V \setminus P_i)|$$

and thus

$$\phi(P_i) \begin{cases} \leq \phi(Z_i) - |\rho_{\mathcal{P}}(k)| \cdot \phi(P_i) & , \text{ if } S_i \neq \emptyset; \\ = \phi(Z_i) & , \text{ otherwise.} \end{cases}$$

However, this contradicts $\Phi(P_1, \dots, P_k) > \rho(k)$ and thus the statement follows. ■

We are now ready to prove Theorem 9.2.

Proof of Theorem 9.2. Let $(P_1, \dots, P_k) \in \mathcal{O}_{\mathcal{P}}$ be a k -way partition compatible with a k -disjoint tuple $Z \in \mathcal{Z}_k$ that satisfies $\Phi(Z_1, \dots, Z_k) = \rho(k)$. By Lemma 9.5 there is a real number such that

$$\kappa_{\mathcal{P}} \triangleq [1 - \rho_{\mathcal{P}}(k)]^{-1} \in (1, k-1]. \quad (54)$$

We argue in a similar manner as in Lemma 9.5 to obtain

$$\phi(P_i) \begin{cases} \leq \phi(Z_i) - \rho_{\mathcal{P}}(k) \cdot \phi(P_i) & , \text{ if } S_i \neq \emptyset; \\ = \phi(Z_i) & , \text{ otherwise.} \end{cases} \quad (55)$$

By combining (54) and the first conclusion of (55) we have

$$\phi(P_i) \leq [1 - \rho_{\mathcal{P}}(k)]^{-1} \cdot \phi(Z_i) = \kappa_{\mathcal{P}} \cdot \phi(Z_i). \quad (56)$$

The statement follows by combining (52) and (56), since

$$\tilde{\rho}_{\text{avr}}(k) \leq \frac{1}{k} \sum_{i=1}^k \phi(P_i) \leq \frac{\kappa_{\mathcal{P}}}{k} \sum_{i=1}^k \phi(Z_i).$$

■

References

- [1] C. Boutsidis and M. Magdon-Ismail. Faster svd-truncated regularized least-squares. In *2014 IEEE International Symposium on Information Theory, Honolulu, HI, USA, June 29 - July 4, 2014*, pages 1321–1325, 2014. doi: 10.1109/ISIT.2014.6875047. URL <http://dx.doi.org/10.1109/ISIT.2014.6875047>.
- [2] C. Boutsidis, P. Kambadur, and A. Gittens. Spectral clustering via the power method - provably. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 40–48, 2015. URL <http://jmlr.org/proceedings/papers/v37/boutsidis15.html>.
- [3] K. R. Davidson and S. J. Szarek. Chapter 8 local operator theory, random matrices and banach spaces. volume 1 of *Handbook of the Geometry of Banach Spaces*, pages 317 – 366. Elsevier Science B.V., 2001. doi: [http://dx.doi.org/10.1016/S1874-5849\(01\)80010-3](http://dx.doi.org/10.1016/S1874-5849(01)80010-3). URL <http://www.sciencedirect.com/science/article/pii/S1874584901800103>.
- [4] G. H. Golub and C. F. Van Loan. *Matrix computations*. Johns Hopkins studies in the mathematical sciences. The Johns Hopkins University Press, Baltimore, London, 2012. ISBN 0-8018-5413-X. URL <http://opac.inria.fr/record=b1103116>.
- [5] J. R. Lee, S. Oveis Gharan, and L. Trevisan. Multi-way spectral partitioning and higher-order cheeger inequalities. In *Proceedings of the Forty-fourth Annual ACM Symposium on Theory of Computing, STOC '12*, pages 1117–1130, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1245-5. doi: 10.1145/2213977.2214078. URL <http://doi.acm.org/10.1145/2213977.2214078>.
- [6] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems 14*, pages 849–856. MIT Press, 2002.
- [7] R. Ostrovsky, Y. Rabani, L. J. Schulman, and C. Swamy. The effectiveness of lloyd-type methods for the k-means problem. *J. ACM*, 59(6):28:1–28:22, Jan. 2013. ISSN 0004-5411. doi: 10.1145/2395116.2395117. URL <http://doi.acm.org/10.1145/2395116.2395117>.
- [8] S. Oveis Gharan and L. Trevisan. Partitioning into expanders. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2014, Portland, Oregon, USA, January 5-7, 2014*, pages 1256–1266, 2014. doi: 10.1137/1.9781611973402.93. URL <http://dx.doi.org/10.1137/1.9781611973402.93>.
- [9] R. Peng, H. Sun, and L. Zanetti. Partitioning well-clustered graphs: Spectral clustering works! In *Proceedings of The 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3-6, 2015*, pages 1423–1455, 2015. URL <http://jmlr.org/proceedings/papers/v40/Peng15.html>. see also <http://arxiv.org/abs/1411.2021>.
- [10] A. Sankar, D. A. Spielman, and S. Teng. Smoothed analysis of the condition numbers and growth factors of matrices. *SIAM J. Matrix Analysis Applications*, 28(2):446–476, 2006. doi: 10.1137/S0895479803436202. URL <http://dx.doi.org/10.1137/S0895479803436202>.
- [11] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000. ISSN 0162-8828. doi: <http://doi.ieeecomputersociety.org/10.1109/34.868688>.

- [12] U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, pages 395–416, 2007.

A Singular Values Bounds of Random Matrices

Lemma A.1 (Norm of a Gaussian Matrix [3]). *Let $M \in \mathbb{R}^{n \times k}$ be a matrix of i.i.d. standard Gaussian random variables, where $n \geq k$. Then, for $t \geq 4$, $\Pr\{\sigma_1(M) \geq t\sqrt{n}\} \geq \exp\{-nt^2/8\}$.*

Lemma A.2 (Invertibility of a Gaussian Matrix [10]). *Let $M \in \mathbb{R}^{n \times n}$ be a matrix of i.i.d. standard Gaussian random variables. Then, for any $\delta \in (0, 1)$, $\Pr\{\sigma_n(M) \leq \delta/(2.35\sqrt{n})\} \leq \delta$.*

Lemma A.3 (Rectangular Gaussian Matrix). *Let $S \in \mathbb{R}^{n \times k}$ be a matrix of i.i.d. standard Gaussian random variables, $V \in \mathbb{R}^{n \times \rho}$ be a matrix with orthonormal columns and $n \geq \rho \geq k$. Then, with probability at least $1 - e^{-2n}$ it holds $\text{rank}(V^T S) = k$.*

Proof. Let $S' \in \mathbb{R}^{n \times \rho}$ be an extension of S such that $S' = [S \ S'']$, where $S'' \in \mathbb{R}^{n \times \rho-k}$ is a matrix of i.i.d. standard Gaussian random variables. Notice that $V^T S' \in \mathbb{R}^{\rho \times \rho}$ is a matrix of i.i.d. standard Gaussian random variables. We apply now Lemma A.2 with $\delta = e^{-2n}$ which yields with probability at least $1 - e^{-2n}$ that $\sigma_\rho(V^T S') > 1/(2.35 \cdot e^{2n} \sqrt{\rho}) > 0$ and thus $\text{rank}(V^T S') = \rho$. In particular, $\text{rank}(V^T S) = k$ with probability at least $1 - e^{-2n}$. ■